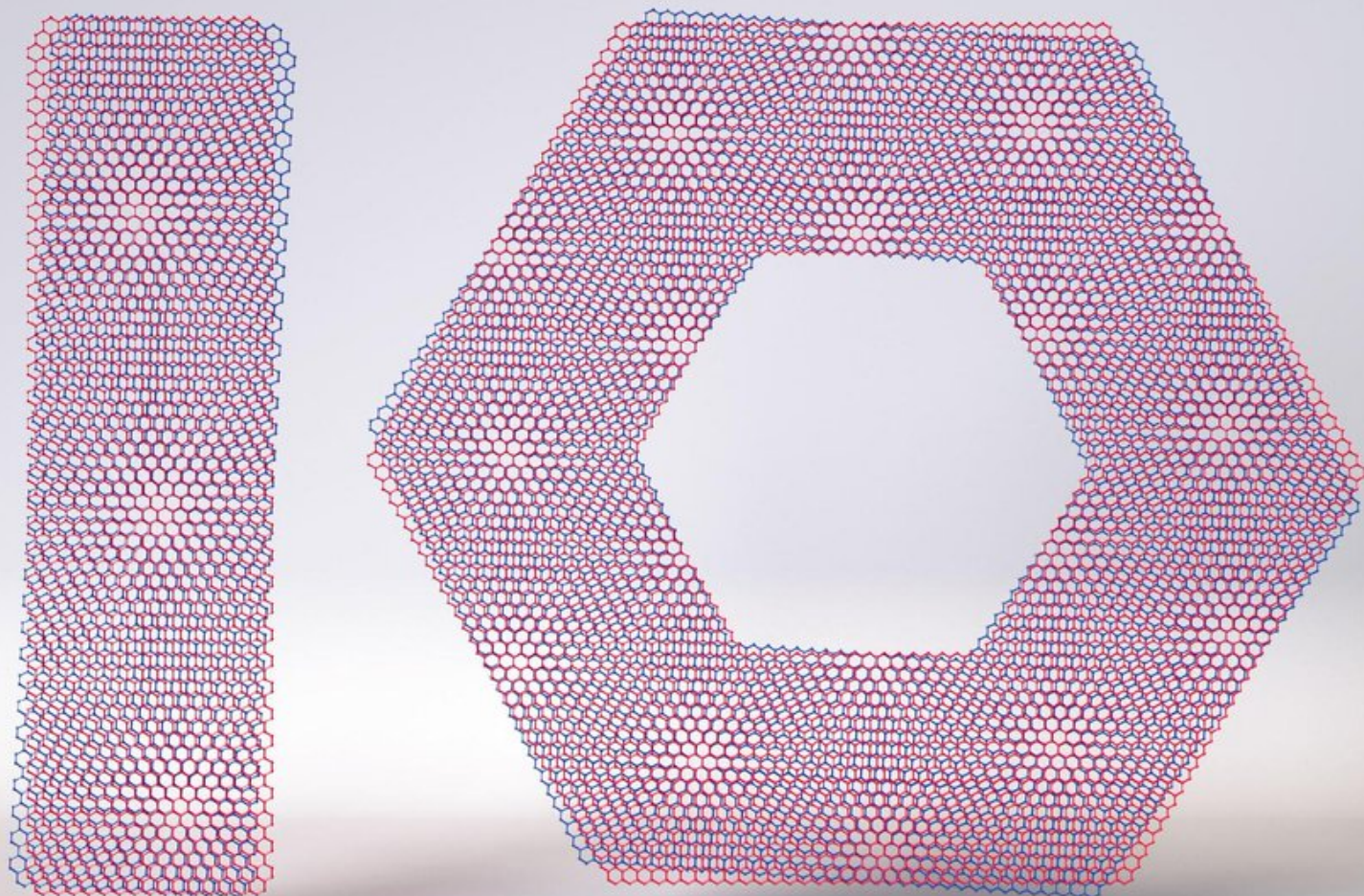


nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



ONE YEAR. TEN STORIES.

Ten people who mattered this year **PAGE 325**

NEWS REVIEW

365 DAYS IN SCIENCE

*From Brexit and wildfires to
the Milky Way and beyond*

PAGE 314

MEDICAL RESEARCH

CROSS-SPECIES TRANSPLANTS

*Improved survival rates for
baboons receiving pig hearts*

PAGES 352 & 430

CALENDAR

EVENTS DIRECTORY 2019

*The Nature guide to global
science events and courses*

BACK PAGES & NATURE.COM

NATURE.COM

20/27 December 2018

Vol. 564, No. 7736

THIS WEEK

EDITORIALS

COMPUTING The future might be quantum — but the classical route is still valid **p.302**

WORLD VIEW One climate scientist's bid to reach 'net zero' emissions **p.303**



MICROBIOME Fruit bat neighbours share fur bacteria **p.304**

Earthrise at 50

An iconic photo of Earth from the Moon was taken by Apollo 8 astronaut William Anders in December 1968. It inspired people then, and can do so again.

It takes an eye for a certain type of detail to look at a photo of the bejewelled Earth hanging in the sky over the sterile terrain of the Moon and see, not the fragility of humanity's only home, but a barren lunar crater. But look carefully and it's there. And now that crater has a new name: Anders' Earthrise.

The Working Group for Planetary System Nomenclature of the International Astronomical Union (IAU) approved the naming of the crater — and another nearby, called 8 Homeward — to mark the 50th anniversary of the Apollo 8 mission that orbited the Moon, and more specifically, the famous photograph taken from on board of Earth rising over the lunar surface. Snapped on 24 December 1968 by astronaut William Anders, *Earthrise* is often labelled as one of the most important and influential photographs in science, if not all of human history.

Needless to say, the status of the image is not down to the circular dent captured in one corner. Instead, it's because the photograph — which seems to show Earth rising above the Moon's horizon — has been credited with starting the environmental movement. Readers of Rachel Carson's book *Silent Spring* — which highlighted the damaging impacts of pesticides on the natural world six years earlier — might argue with that common trope. But it's undeniable that *Earthrise* was profoundly important in raising awareness and focusing minds. For the first time, people could see their planet framed against the black emptiness of eternal space and appreciate its technical colour beauty as well as its utter insignificance in the Universe.

An entire generation suddenly saw the planet as isolated and vulnerable, and very difficult to replace. (A later generation would experience this for themselves, with the publication of another iconic picture of the planet: the *Pale Blue Dot*, taken from a distance of 6 billion kilometres by the Voyager 1 probe on St Valentine's Day in 1990.)

The view of Earth from space is much the same now as it was then. (Just witness the stunning images released earlier this year from the GOES-16 satellite, which show the planet in extraordinary detail.) But how we think about such images has drastically changed.

For many millions of people, the end of 2018 sees a better, more prosperous world than the one the Apollo 8 astronauts returned home to 50 years ago. Human progress, driven by advances in science, medicine and technology, has radically improved average living standards, health and life expectancy. But Earth itself is panting to keep up. Only two months ago, the Intergovernmental Panel on Climate Change issued its most urgent warning yet about the effects of climate change, warning that a temperature rise of even 1.5°C — which most experts agree is inevitable — will bring devastating droughts and floods.

It's likely to be much worse than that, however. Last week, the world's politicians met in Poland to discuss next steps on a global climate

agreement that could be the last, best hope to stem climate change. The deal will make insufficient change to the amounts of damaging greenhouse-gas emissions we hurl into the atmosphere.

Powerful images show what is at stake. But they also show what we can still achieve: that we do not have to be passive observers, trapped by the scale and magnitude of the Universe and its problems. We can act. We can make things happen.

Take *Earthrise*, the picture and the phenomenon. We did that. The Moon is tidally locked to Earth and that fixes the planet's position in the lunar sky. Earth doesn't rise from the Moon and only seemed to do so for the Apollo 8 astronauts because their craft was speeding above the surface, gradually revealing more of Earth as it travelled. Even as the planet hung there in the blackness of infinity, the people who saw it were moving forwards. We still can. ■

Fur and fossils

Feather-like structures on pterosaurs open up a world of colour.

Pterosaurs are the first known vertebrate group to have evolved powered flight — preceding birds and bats by many millions of years. Ranging from the size of small birds to that of small planes, pterosaurs lived alongside the dinosaurs and went extinct at the same time. Many things about these creatures remain mysterious, not least their origin — the earliest pterosaur fossils found so far seem to have been fully capable of flight, and there is no confirmed transitional fossil to show from which reptilian group they emerged.

This is different from, say, birds. Revelations over the past two decades that bird-like feathers were present on dinosaurs — ground dwelling and with the flight capability of a sack of spanners — have illuminated our understanding of the evolution of birds and their characteristic structures.

That the bodies of at least some pterosaurs were clothed with a kind of fuzz has been known (or at least suspected) since the 1830s, but this fluffiness became a focus of study only after the description of the exceptionally hirsute Kazakh pterosaur *Sordes pilosus* in 1971.

Pterosaur fluff, comprised of what are technically known as 'pynofibres', is structurally different from mammalian fur or hair. Each pynofibre is a short, simple filament with a canal running down the centre, and is much more superficially attached than the deeply rooted hairs of mammals. Pynofibres have been observed on the heads, limbs and bodies of several pterosaur fossils.

Ironically, given that they could fly, discussion of feathers and

feather-like structures has tended to ignore pterosaurs. Instead it has focused on non-avian dinosaurs, which couldn't. As a result, the relationship — if any — between pterosaur pycnofibres and dinosaur feathers has been obscure.

No longer. A paper this week in *Nature Ecology and Evolution* shows that some pycnofibres, far from being simple monofilaments, had branching or brush-like structures — just like the feathers found on birds and their closest dinosaur relatives (Z. Yang *et al.* *Nature Ecol. Evol.* 3, 24–30; 2019).

The study suggests, therefore, that pycnofibres could share an evolutionary origin with dinosaur and bird feathers. And the common ancestor of birds, dinosaurs and pterosaurs might also have been able to produce such pycnofibre structures.

The study's evidence comes from fossils of two sparrow-sized pterosaurs between 160 million and 165 million years old (a shade earlier than the earliest known bird, *Archaeopteryx*, which is around 150 million years old), from the Jurassic period of China.

The pterosaurs have four distinct kinds of pycnofibre: the regular monofilaments seen in other pterosaurs; a type with a brush at the distal ends; a variety in which brush-like filaments sprout from the middle of the main fibre; and a fourth, in which several fibres meander from a common root. Structures corresponding to all four types of pycnofibre have been found associated with various dinosaurs, underlining the case that pterosaurs are indeed related to dinosaurs.

Importantly, each kind of fibre is not distributed randomly on the bodies of the two pterosaurs. The simple monofilament form is found all over the body; the brush-like form on particular regions of the head, limbs and tail; and the curious form with sprouting filaments is

restricted to the head. The fourth form, which closely resembles the down of bird chicks, is found on the wing membranes.

This distinct distribution indicates that each type had a biological function, and that one kind of filament was not simply the decayed product of another.

What were these functions? Pycnofibres of the first and second type might have provided insulation and streamlined the body shape

“For the first time, we can visualize pterosaurs with a touch of colour.”

to minimize aerodynamic drag, as feathers do in birds and fur does in bats. The sprouting type on the head might have functioned similarly to the sensory bristles found on the heads of modern birds. The downy, fourth kind of fibre might have helped to keep the wings warm, as it's known that feathers with this structure are much more efficient at trap-

ping warm air than is mammalian hair.

Moreover, the pycnofibres contain remnants of melanosomes — organelles that are typically found in feathers, feather-like structures and mammalian hairs, and that help to lend these structures their distinctive colours. When applied to the pterosaur fuzz, a technique called Fourier-transform infrared spectroscopy produces the same spectra as those found in birds both ancient and modern, as well as red (but not black) human hair.

For the first time, we can visualize pterosaurs with a touch of colour, as we can fossil birds, dinosaurs and even dinosaur eggs. Flying alongside the earliest birds and even some very early flying mammals, pterosaurs must have made the skies of the Mesozoic Era a riot of life and colour. ■

Computer games

Classical and quantum machines are battling for computational superiority.

Will 2019 be the year when quantum computers show they have the right stuff? Google says so — one of the company's labs, in Santa Barbara, California, has promised that its state-of-the-art quantum chip will be the first to perform calculations beyond even the best existing supercomputers.

And Google isn't alone. A number of other companies, big and small, are working steadily towards the same symbolic goal. Venture capitalists have poured money into dozens of quantum-computing start-up companies. Excitement and anticipation are mounting.

In a stark reminder of the power of quantum computing, in May, two theoretical computer scientists solved a 25-year-old conjecture (go.nature.com/2eatyco). They confirmed that quantum computers are — in an admittedly abstract setting — vastly more efficient than classical ones at particularly complex tasks, such as testing whether a set of numbers is random.

Still, such work does not justify the expectations that now surround quantum computing. A recent report by the US National Academies of Sciences, Engineering, and Medicine (penned by leading Google and Microsoft researchers, among others) stressed the technical hurdles that lie in the way of building practically useful quantum computers. Creating such machines will take at least a decade, the report says.

Theoretical physicist Seth Lloyd at the Massachusetts Institute of Technology in Cambridge speaks for many when he says the field is in a period of explosive progress — but that the hype is also getting out of control. “The whole quantum-computing field is just going hogwild right now,” he says.

Is a quantum computer even needed? High-profile work by an 18-year-old computer scientist earlier this year suggests not, at least for one specific task. Ewin Tang effectively taught an old computer a new

trick — one that was previously thought to need a quantum system.

She developed an extremely efficient classical algorithm — that is, one that can run on an ordinary computer — for ‘recommendation systems’, such as those that certain websites use to try to guess a consumer's tastes (E. Tang Preprint at <https://arxiv.org/abs/1807.04271>; 2018). Her work produced a much faster version than current, relatively sluggish systems. Tang's algorithm is not necessarily practical to use, so it won't replace current algorithms unless it is substantially improved — in its current form, it would be useful only with data sets of truly gigantic proportions. But a quantum algorithm that was in development for that same task has now been rendered moot, before it ever had a chance to run on an actual machine.

Last month, Tang, who is now at the University of Washington in Seattle, doubled down. She and two colleagues demolished the quantum advantage of another type of algorithm for certain machine-learning tasks (A. Gilyén *et al.* Preprint at <https://arxiv.org/abs/1811.04909>; 2018). A different team at the University of Texas in Austin reached the same conclusion independently (N.-H. Chia *et al.* Preprint at <https://arxiv.org/abs/1811.04852>; 2018). Computer scientists responded to the news with memes that, for example, compared Tang to a gladiator slaughtering the hopes and dreams of the quantum community. And it was a bittersweet moment for Tang's co-author, Seth Lloyd — he wrote the quantum algorithm that was trounced.

Some in the field argue that these uses of classical computing are actually successes for quantum computing, because they show how the quantum way of thinking can have an impact, even before quantum computers exist. Specialists also point to problems for which quantum computers have long been known to have a proven advantage, such as web searches. In other cases — such as factoring large integers into primes or simulating the electronic properties of materials — scientists think that quantum computers are still likely to have an advantage, although this has not yet been demonstrated mathematically.

Quantum computers are a not-yet-existent technology in search of problems to solve. Meanwhile, researchers are seeing how far classical strategies can be taken. Both are valid research avenues. A quantum device remains a laudable goal. But it's not the only route to the future. ■



How I stave off despair as a climate scientist

So much warming, so many dire effects, so little action — Dave Reay reveals how dreams of soggy soil and seaweed keep him going.

There's a curve that is quietly plotting our performance as a species. This curve is not a commodity price or a technology index. It has no agenda or steering committee. It is the Keeling curve. It is painfully consistent in its trajectory and brutally honest in its graphical indictment of our society as one that stands ready to stand by as islands submerge, cities burn and coasts flood.

Established by Charles David Keeling in 1958, the curve records how much carbon dioxide is in our atmosphere — fewer than 330 parts per million then, more than 400 today. Each month for the past decade, my geeky addiction has been to scan the latest data. To search for some hint that 'Stabilization Day' will come: when global emissions and global uptake are once more in balance. As yet another 'last-chance' United Nations climate-change meeting draws to a close, emissions are still rising.

In climate science, you can check out of the lab anytime you like, but you can never leave. The overheating Earth that our super-computers model is the one we all share and which our children will inherit. Dynamic, high-resolution representations of warming trends and weather patterns that delight me as a researcher chill my spine as a human being: I stare at the lines curving up and see the people who endure them.

There are days when refining another obscure step on the road towards climate catastrophe gets you down. Some colleagues reach for gallows humour to keep them going — the quip "we're going to need a bigger boat" is common in the face of the latest damning assessment of global inaction. Others seek solace in uncertainty, grasping at the coolest strands of future projections: the green pathways of a rapid and sustained global response. Many of us — my younger self included, as I expounded in my book *Climate Change Begins at Home* — try to wrestle back an iota of control by cutting our personal carbon footprints and spending our salaries on solar panels, super-insulated homes and electric cars. A few of us have foregone air travel and openly questioned how those who work on climate change can justify a high-emissions lifestyle.

Every tonne of carbon emissions avoided does matter, but unless individual actions are replicated globally, we are pissing in a hurricane. By the middle of this century, the world must reach net zero emissions. So, what more is an academic to do? Write more *Nature* papers? Blockade the university car park? Knit our elbow patches from hemp?

For me, the most powerful response is to teach. By educating new waves of practitioners, policymakers and researchers, I can vicariously boost mitigation and adaptation capacity at scales and across time horizons I could never reach alone. On restless nights, when futures of famine and storm-surge devastation play out

behind my eyelids, that's what helps me sleep.

That, and a personal plot to pull a lifetime's worth of carbon out of the atmosphere.

The dream with which I've bored my family to distraction for the past 20 years is going truly 'net zero': paring down emissions to the bare minimum, and then managing a chunk of land to try to sequester the remainder.

Last month, that dream came true. Years of saving, a large dollop of luck and an even larger loan made me and my wife the nervous owners of 28 hectares of rough grassland and wild rocky shores in the west of Scotland. The coming years will see us map every baseline carbon stock and flux, from the soil and vegetation, to the bemused

sheep and 'blue carbon' of the seaweed beds. Each gnarled tree trunk will be hugged with a tape measure, every soggy field corner will be probed, sampled and analysed. We'll then plant trees. Lots of them — native tree species that will boost biodiversity, draw down carbon dioxide and withstand the inevitably turbulent decades and centuries to come.

As a research project, it is a chance to verify the science, and test the concepts of climate-smart land use in the teeth of Atlantic storms and hungry deer. As our future home, it is the chance to finish life as we started it: with an atmospheric blank slate.

Of course, this dream of sustainability is not itself sustainable. My family and I are fortunate to be well-off people in a rich country. To replicate this for every person in the world would require many, many times the area of land that is actually available. We are embarking on a

privileged journey that billions could never hope to take, and that, even at its emissions-trapping best, will hardly register in Scotland's national carbon account. Hopefully, my students can magnify its impact — learn from our trials and errors and help to take such carbon-management expertise global.

We've long known that reaching 'net zero' globally will require our emissions to plummet, but that some emissions are unavoidable. Worldwide, this will necessitate large increases in tree planting, soil enhancement and other such carbon-capture strategies.

The Keeling curve might remain a monthly glimpse into the abyss, but alongside it will now be a personal emissions curve that holds a real possibility of hitting the x-axis. Field trips for my climate-change classes are about to get a whole lot more hands-on. As a carbon geek, I've never been so excited to take my work home with me. ■

Dave Reay is a professor of carbon management and education at the University of Edinburgh, UK.
e-mail: david.reay@ed.ac.uk

REPRESENTATIONS OF
WARMING
TRENDS THAT
DELIGHT ME
AS A RESEARCHER
CHILL MY
SPINE
AS A HUMAN
BEING.

SEVEN DAYS

The news in brief

EVENTS

Students blocked

Two more universities in Japan have admitted to systematically favouring male applicants to their medical degrees over women. On 10 December, Juntendo University and Kitasato University, both in Tokyo, posted statements on their websites acknowledging the practice. The revelations come four months after reports that Tokyo Medical University had been altering entrance-examination results for years to keep the proportion of female entrants below 30% of all students. A government investigation into whether the practice was used at other medical schools found that Juntendo University had unfairly assessed 164 applicants in 2017 and 2018. Of those, 117 who had failed medical-school exams on previous application attempts, including 74 women, were unfairly blocked from proceeding to the second stage of assessment. Another 47 women and one man who should have been accepted were denied entry. The university's statement said it was compensating for the difference in emotional maturity between men and women at the age of college entrance. Kitasato University said it has launched an independent investigation into its own actions. Both universities have vowed to end the practice in 2019.

POLICY

Climate talks

Delegates from nearly 200 countries have come to an agreement on how to implement the 2015 Paris climate accord. The deal, reached on 15 December at the United Nations climate summit in Katowice, Poland, establishes rules for tracking and reporting greenhouse-gas

emissions and climate policies. It also lays out processes intended to boost national efforts to curb emissions and increase the transfer of money and technologies to developing countries. "The multilateral system has delivered a solid result," UN climate chief Patricia Espinosa said in a statement. "This is a roadmap for the international community to decisively address climate change."

EU science scheme

The European Parliament has adopted proposals for Horizon Europe, the European Union's next big research-funding programme, for 2021–27. The parliament proposed altering the way

in which non-EU member states can participate in the scheme, which could benefit the United Kingdom after Brexit. One proposed category would include countries that are closely associated with the EU, such as Switzerland and Norway. A second tier would be for nations with looser ties. Britain's category would depend on the access deal agreed with the European Commission after Brexit, said one of the parliament's rapporteurs on Horizon Europe. But Britain will continue to be treated as a member state in negotiations until it leaves the EU. The parliament also voted to raise the scheme's budget from nearly €100

billion (US\$114 billion) to around €120 billion. The proposals must now go before the Council of the European Union.

Gene-editing rules

An international agreement on standards for germline gene editing is needed urgently, said the presidents of the Chinese Academy of Sciences and the US National Academies of Sciences, Engineering, and Medicine in a 13 December essay (V. J. Dzau *et al. Science* 362, 1215; 2018). The call is in response to a widely condemned claim that a Chinese scientist used the CRISPR–Cas9 tool to edit genes in two human embryos, leading to the birth of twin



KATHRYN HANSEN/NASA

Report flags trouble for the Arctic

The Arctic experienced its second-warmest year on record between October 2017 and September 2018, according to a report released by the US National Oceanic and Atmospheric Administration on 11 December. Average air temperatures in this frigid region have hit record or near-record levels every year since 2014. And rising temperatures have

contributed to more than a 50% decline in wild reindeer and caribou populations since the 1990s. Government researchers also reported that roughly 99% of sea ice is now considered relatively new, meaning that it hasn't lasted for more than four summers without melting. It's now the thinnest and most susceptible to warming temperatures that it's been in 30 years.

girls last month. The essay calls for an expedited report from international science academies, which advise governments. It cites as a model the 1975 Asilomar Conference on Recombinant DNA, in which scientists established voluntary guidelines for safely working on the technology. Some researchers, however, contend that stricter, legally enforceable limits are needed on germline gene editing.

Science advice

Spain is set to get an official science advisory panel that will gather scientific evidence on a range of technical and social issues for its lawmakers. The parliament's lower house, the congress, designated €200,000 (US\$227,000) of its draft 2019 budget on 3 December to creating an office of science and technology. It will be modelled on the UK Parliamentary Office of Science and Technology (POST), which provides evidence at early stages of political debates on scientific issues. The Spanish office will take a broader approach to science, also covering issues such as immigration and gender equality. The move follows a November meeting between parliamentarians (pictured, Prime Minister



Pedro Sanchez) and Science in the Parliament (Ciencia en el Parlamento), a grass-roots organization of scientists that has campaigned for such a body since late 2017. Ciencia has also acted as a voluntary, unofficial POST in Spain over the past few months, commissioning reports on 12 topics picked by parliamentarians from 50 themes that emerged from a wider public consultation.

PEOPLE

Antarctic deaths

Two fire technicians at the US National Science Foundation's (NSF) McMurdo Station in Antarctica were pronounced dead on 12 December after an incident at a generator building. A helicopter pilot noticed what appeared to be smoke rising from the building and found the two people lying unconscious inside. It's unclear what happened and the NSF

says that an investigation is under way. The individuals, whom the NSF would not identify, had been testing the operation of the structure's fire-suppression system. Medical personnel from the McMurdo clinic declared one of the workers dead at the scene. The other worker was flown to the McMurdo clinic and pronounced dead shortly thereafter.

FUNDING

NIH crackdown

The US National Institutes of Health (NIH) is cracking down on private donations to research projects, following revelations earlier this year that researchers studying drinking had solicited funding from the alcohol industry in violation of NIH rules. The agency will review all projects receiving private funds to identify potential conflicts of interest, said deputy director Lawrence Tabak at a 13 December meeting of the agency's Advisory Committee to the Director (ACD) in Bethesda, Maryland. In an effort to prevent NIH programme officers from favouring certain companies, the agency will also begin publicly disclosing more information about unsuccessful applications for partnerships, as well as how

private funds are awarded. The NIH terminated the alcohol study in June after the ACD decided that its results could not be trusted.

FACILITIES

Fusion energy

The United States should continue its participation in the international fusion-power project ITER while expanding its domestic research and development efforts in the field, says a report released on 13 December by the National Academies of Sciences, Engineering, and Medicine. The troubled multibillion-euro project has faced scepticism from members of the US Congress, but the United States has continued to contribute to the project. The country's current budget includes US\$132 million for ITER, an increase of \$10 million from the previous year. To capitalize on that investment — and avoid being overtaken by other countries — the United States must also expand its domestic efforts to advance fusion energy, the report says. It recommends that the country increase its spending on fusion research by nearly \$200 million annually for the next several decades, so that it can maintain its partnership in ITER and build its own pilot plant.

TREND WATCH

Global spending on tuberculosis (TB) research and development hit a high in 2017, according to a report from the activist organization Treatment Action Group (TAG) in New York City.

Investment reached US\$772 million, up from \$726 million in 2016. The hike follows two years of declines, in 2014 and 2015. It is the most spent on research into TB in a year, but still falls short of the \$2 billion a year that the TB research community says is needed to end the disease by 2030. That target is one of the United Nations Sustainable Development Goals and part of

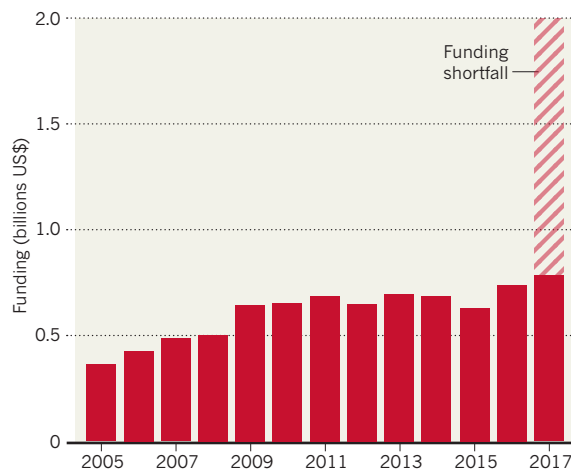
the World Health Organization Global Plan to End TB.

Around 25% of the world's population — 1.8 billion people — is infected with TB, the World Health Organization estimates. In 2017 alone, some 10 million people fell ill with the disease, and 1.6 million died.

The disease remains prevalent but science is generating hope that new and improved means of diagnosis, vaccination and treatment are on the horizon. "We're at an incredibly promising moment in TB research globally," says Mike Frick, TB project co-director for TAG.

TUBERCULOSIS FUNDING SHORTFALL

Research and development funding for tuberculosis is on the rise, but it is still well below the US\$2 billion a year needed to end the disease by 2030.



NEWS IN FOCUS

PUBLIC HEALTH Italian scientists oppose donation for vaccine-safety research **p.310**

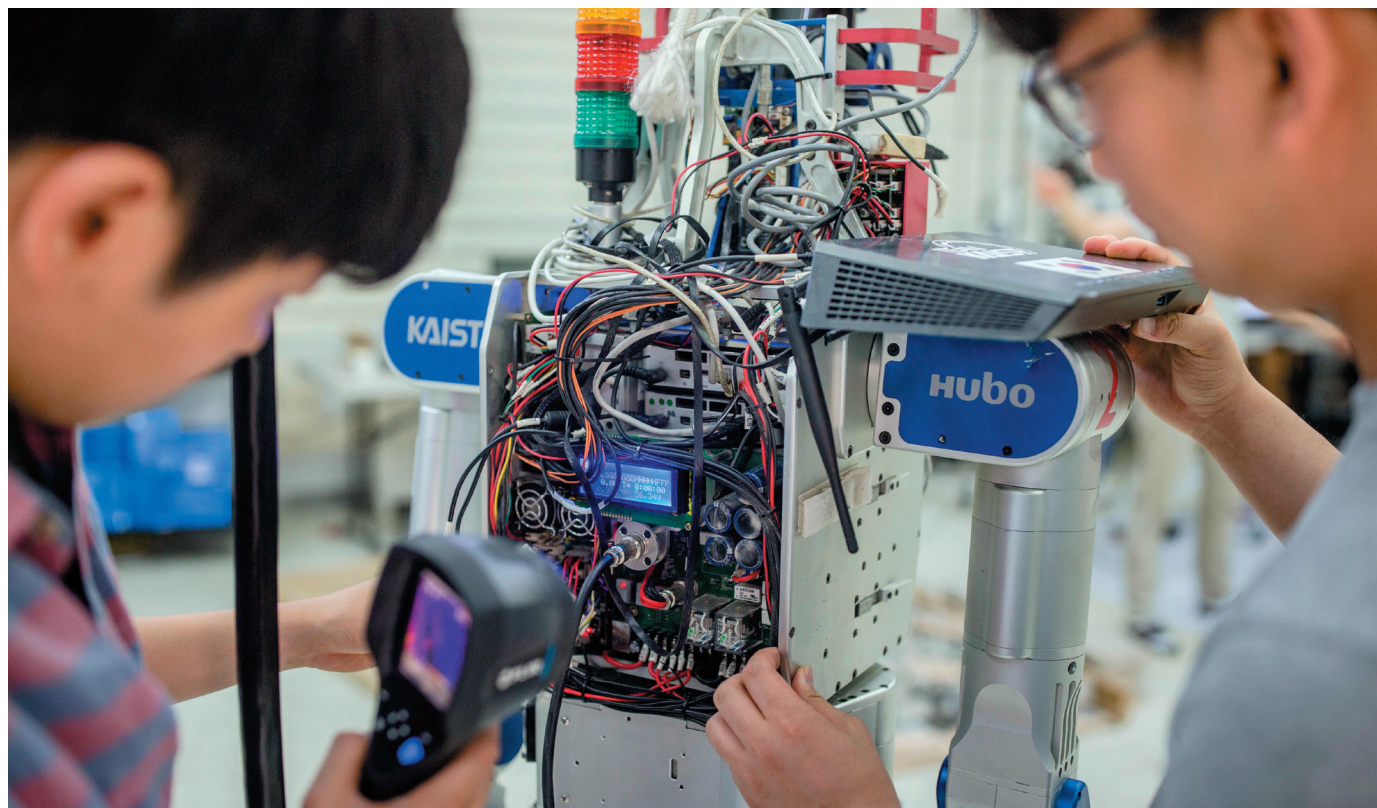
CLIMATE Southern Ocean spotted burping carbon dioxide on the sly **p.311**

CHINA Penalties for research misconduct go beyond academic career **p.312**



NATURE'S 10 Ten people who mattered in science in 2018 **p.325**

JONAS GRATZER/LIGHTROCKET VIA GETTY



Leaders at South Korea's top technical university, KAIST, have voted to delay a decision to suspend its president.

SOUTH KOREA

Outcry over treatment of university president

Government alleges Shin Sung-Chul misused funds, but scientists see a political purge.

BY MARK ZASTROW

Researchers in South Korea have criticized the nation's science ministry for its handling of an investigation into the president of the Korea Advanced Institute of Science and Technology (KAIST) in Daejeon.

The ministry alleges that Shin Sung-Chul misused public funds in his previous job by making payments to Lawrence Berkeley National Laboratory (LBNL) in California, accusations that Shin denies. The payments

were part of a deal in which scientists at South Korean universities were guaranteed access to one of LBNL's imaging facilities, an X-ray microscopy beamline. The ministry has referred the allegations to prosecutors, and requested that KAIST's board of trustees suspend Shin — a decision that the board deferred when it met on 14 December.

Many scientists suspect that the allegations are part of a politically motivated attempt to remove Shin, who was hired under the administration of South Korea's previous president.

The call to suspend him seems to have been rushed and is based on insufficient evidence, they say. Researchers also say that the ministry has misconstrued a common practice in which institutions pay fees to use equipment at international facilities.

The ministry says that the payments — 2.2 billion won (US\$1.9 million) in total — were illegal because they were not part of a 2012 agreement between LBNL and the Daegu Gyeongbuk Institute of Science and Technology (DGIST), a publicly ▶

► funded institute where Shin was president from 2011 to 2017 before taking over the presidency of KAIST.

The ministry's audit team uncovered the alleged misuse, and the South Korean broadcaster SBS first publicized the allegations on 25 November. The ministry says that some of the 2.2 billion won was paid to one individual, a former student of Shin's, and that this could constitute embezzlement.

Shin refutes the allegations. In an e-mail from KAIST, sent in response to questions from *Nature's* news team, Shin said that neither he nor DGIST was involved in any illegal activities, misconduct or embezzlement regarding LBNL, which is owned by the US Department of Energy (DOE). "The collaboration contract between two institutions was fully approved through all proper rules and regulations of the DOE and LBNL's contracting processes," he said.

As president of DGIST, Shin brokered an agreement with LBNL in 2012 that gave his institute's researchers the chance to work with one of the world's most respected physics labs. The collaboration received ten weeks of beam time, which DGIST did not pay for. The agreement expanded in 2014, and again two years later, securing DGIST half of all available time on one of LBNL's X-ray microscopy beamlines,

which its researchers used to explore nanomaterials. As part of the arrangement, DGIST paid an annual facility fee.

On 10 December, LBNL sent a letter to the ministry — seen by *Nature* — supporting Shin's version of events. It says that the agreement with DGIST was a customary approach to conducting collaborative research with international partners and that the reported allegations "contain significant errors in fact and in assumptions". The letter also said that the collaboration required significant instrument time beyond the scope of a standard short-term project, and that the payments supported the operation and staffing costs of running the beamline.

A petition in support of Shin, organized by the KAIST physics department — where Shin worked from 1989 to 2011 — had collected more than 830 signatures from researchers at South Korean institutions by 14 December.

The petition says that calls to suspend Shin lack due process because they are based on an ongoing investigation and unproven accusations. "There simply is not enough evidence

"There simply is not enough evidence to justify suspending him from his duties."

allegations "contain significant errors in fact and in assumptions". The letter also said that the collaboration required significant instrument time beyond the scope of a standard

to justify suspending him from his duties," the petition states. The ministry has "treated him like a criminal", says one of the petition organizers, who requested anonymity because they fear retaliation from the government for speaking out.

The science ministry said in two statements that the request for Shin's suspension was carried out in accordance with the ministry's authority to regulate public institutions.

The ministry has also accused Shin, two other DGIST professors and Shin's former student — now a staff scientist at LBNL — of misconduct. It alleges that they did not follow the correct process when granting the former student an adjunct position at DGIST during Shin's presidency, and referred them to prosecutors on 28 November.

Shin told *Nature* that he did not offer the student favourable treatment. In a press conference on 4 December, he also said he had had nothing to do with determining their salary or hiring at LBNL or DGIST.

In its letter, LBNL says its researcher is an expert in soft X-ray microscopy; that it had followed its own hiring and salary disbursement procedures; and that no DGIST funds had been sent directly to the researcher.

The ministry did not respond to *Nature's* questions about LBNL's letter or the petition. ■

ITALY

Scientists slam donation to question vaccine safety

Italian National Order of Biologists donated €10,000 for research into vaccine ingredients.

BY GIORGIA GUGLIELMI

Some scientists in Italy are up in arms over a donation from the organization that oversees the nation's professional biology qualification to an advocacy group that opposes mandatory childhood vaccination.

The news comes as Italian politicians debate whether to continue with the mandatory vaccination policy, which was introduced in 2017 and requires parents to provide proof of ten routine vaccinations when enrolling their children in nurseries and preschools.

The advocacy group, Corvelva, announced that it had received €10,000 (US\$11,350) from the National Order of Biologists (ONB) on 26 October says that it plans to use the money for research that investigates the safety and efficacy of commonly used vaccines. Corvelva says that previous studies it has funded, which have not yet been published in a peer-reviewed

journal, indicate that some vaccines contain impurities, or lack the active ingredients they are claimed to contain.

ONB president Vincenzo D'Anna told *Nature* in an e-mail that there is a need for truly independent vaccine research because,

"Studies that monitor reactions 'cannot exclude the possibility that vaccines are toxic.'"

in his opinion, work conducted in public laboratories and at universities is usually influenced or funded by companies that produce vaccines.

"The goal is to contribute to complete the biological and chemical analyses on vaccines," he said in the e-mail interview, part of which the ONB has published in its Bulletin.

But many scientists dismiss the need for the additional research — on the grounds that vaccines are already rigorously tested — and are

flummoxed by the ONB's donation.

"There's solid evidence that vaccines work and are safe," says virologist Giorgio Palù at the University of Padova, who is president of the European and Italian societies for virology.

Membership in the ONB confers certification for jobs in the biological sciences in Italy. The order has about 50,000 members who each pay an annual membership fee of €120.

The large-scale, expensive studies that test vaccines' efficacy and monitor for adverse side effects are regulated and supervised by national and international health agencies and are "far more accurate than tests that could be done with €10,000", says Gennaro Ciliberto, a molecular biologist at the University of Catanzaro Magna Graecia and president of the Italian Federation for the Life Sciences, which includes 14 scientific societies.

Once vaccines are approved, these agencies continue to monitor them by testing batches

and production facilities for safety, as well as tracking adverse reactions, he adds.

But Marchi says studies that monitor adverse reactions don't track participants for long enough, and "cannot exclude the possibility that vaccines are toxic".

D'Anna emphasized that the donation to Corvelva is not the full amount that will be spent on the research.

Corvelva has collected more than €50,000 so far, says Marchi. The organization will use the money to check whether vaccine components are indeed those indicated on the label, and to look for contaminants. Marchi says that the group hopes to influence the debate on whether to continue with the 2017 mandatory-vaccination policy.

Giovanni Maga, a molecular biologist at the National Research Council's Institute of Molecular Genetics in Pavia, worries that the ONB's decision to fund this research could increase public distrust of vaccines.

D'Anna rejected this idea. On the contrary, he said, more people will choose to vaccinate their kids if "we could guarantee them the absolute safety of vaccines".

D'Anna said that neither he nor the ONB can be defined as 'no-vax', a term used in Italy to refer to people who are against vaccinations, and says that he has never questioned the efficacy of vaccines. "The ONB and the biologists know well the merits of vaccines, and want to know all the rest about their safety," he said.

The debate about the donation follows criticism of a conference to celebrate the ONB's 50th anniversary in March. Some academics and scientific societies urged the ONB to revise the agenda because they were concerned that anti-vaccine ideas could be promoted, although the ONB rejects this criticism.

The donation and the choice of speakers at the March meeting are included in a petition calling for the Ministry of Health, which oversees the governance of the ONB, to remove D'Anna as ONB president. The petition, created by three graduate biology students, says that these and other actions by the ONB endanger public health and discredit the scientific community.

In a telephone interview with *Nature*, D'Anna said he won't step down. And in the e-mail interview, he dismissed the seriousness of a petition launched by students. He said that those who want to verify whether "hundreds of biological and chemical impurities" can harm children do not endanger public health.

A spokesperson for the Italian Ministry of Health says that it has received "a report on the matter" of the ONB donation to Corvelva, and that the ministry asked the ONB "to provide information on the subject". The ministry doesn't fund the ONB, but it is tasked with ensuring that the governing board abides by its duties. ■



Icy southern waters help to blunt climate change by pulling carbon dioxide from the atmosphere.

CLIMATE CHANGE

Southern Ocean spotted burping CO₂

Ocean-float data reveal that waters off Antarctica don't absorb as much carbon as scientists thought.

BY JEFF TOLLEFSON

The Southern Ocean is one of humanity's allies, slowing global warming by absorbing heat and carbon dioxide from the atmosphere. But now researchers report that the choppy waters around Antarctica are also quietly belching out massive quantities of CO₂ during the dark and windy winter, reducing the ocean's climate benefit.

The scientists behind the work, presented last week at a meeting of the American Geophysical Union in Washington DC, say that the winter emissions reduce the Southern Ocean's net uptake of CO₂ by 34%, or more than 1.4 billion tonnes per year. That amount is roughly equal to Japan's annual carbon emissions.

"The Southern Ocean is still going to be important in the global carbon cycle," says Seth Bushinsky, an oceanographer at Princeton University in New Jersey who is leading the study. "We're just trying to understand exactly how and why."

The ocean's winter CO₂ emissions, which were tracked by a fleet of robotic floats, occur when deep waters rise to the surface and release centuries-old carbon. This is part of a larger process of ocean circulation that

moves heat and nutrients around the globe, but researchers have struggled to pin down precisely how the overall system works, in part because of a dearth of data.

For years, scientists have based their estimates of carbon uptake in the Southern Ocean on measurements made by ships sailing to and around Antarctica, but the data are sparse — particularly for the winter months.

The latest work factors in 3.5 years of data from 65 floats deployed as part of the US\$21-million Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) project. The floats bob up and down in the upper 2,000 metres of the ocean, measuring temperature, salinity, oxygen, carbon and nutrients — information that can be used to infer how much carbon is moving into and out of the ocean.

The first estimate based on the SOCCOM floats alone, published in August, reduced the Southern Ocean's carbon uptake by more than 90%, compared with previous calculations based on ships' measurements (A. R. Gray *et al. Geophys. Res. Lett.* **45**, 9049–9057; 2018). But the discrepancy raised eyebrows, and prompted concerns about potential bias in the float estimates.

To produce its latest estimate, which ▶

GRAEME SNOW/ALAMY

► includes float and ship data, the SOCCOM team worked with researchers who have helped to produce ship-based estimates for the Global Carbon Project, an international consortium that tracks the carbon cycle. Bushinsky says the team is reviewing incoming float data to discover what is causing the discrepancy with the ship-based measurements. The scientists are also trying to determine whether the ocean's winter CO₂ burps are a regular occurrence or a short-term trend caused by natural variations in ocean circulation.

But the SOCCOM floats have already given scientists a better look at the Southern Ocean, says Corinne Le Quéré, director of the Tyndall Centre for Climate Change Research

in Norwich, UK.

And although the Southern Ocean is doing less to moderate global warming than scientists thought, evidence is mounting that its influence on climate will grow during this century, says Joellen Russell, an ocean modeller at the University of Arizona in Tucson who heads SOCCOM's modelling team.

One recent study found that Antarctic meltwater flowing into the ocean creates a layer of cold, fresh water that pushes warmer, saltier water up under the continent's ice shelves, accelerating ice loss (B. Bronselaer *et al.* *Nature* **564**, 53–58; 2018). Researchers used a climate model to look ahead to 2100, and project that this flow of meltwater will cool the region

and slow the increase in average global temperatures, even as it accelerates Antarctica's contribution to sea-level rise.

Now the SOCCOM team is taking a closer look at the powerful winds that circle Antarctica, which have strengthened and moved polewards over the past several decades. Few climate models simulate this process. But in unpublished research, the SOCCOM scientists have found that their model better reproduces the data collected by floats when it incorporates more-realistic simulations of Antarctic wind patterns and meltwater flows. "We now have a cookie-cutter approach for telling whether our models are getting the Southern Ocean right," Russell says. ■

CHINA

Social punishments for scientific misconduct

Offending researchers could face restrictions on jobs, loans and business opportunities.

BY DAVID CYRANOSKI

Chinese researchers who commit scientific misconduct could soon be prevented from getting a bank loan, running a company or applying for a public-service job. The government has announced an extensive punishment system that could have significant consequences for offenders — far beyond their academic careers.

Under the policy, dozens of government agencies will have the power to hand out penalties to those caught committing major scientific misconduct, a role that was previously the preserve of the science ministry or universities. In addition to existing misconduct penalties, such as the loss of grants and awards, errant researchers could face punishments that have nothing to do with research, including restrictions on jobs outside academia.

"Almost all aspects of daily life for the guilty scientists could be affected," says Chen Bikun, who studies scientific evaluation systems at Nanjing University of Science and Technology.

The policy, announced last month, is an extension of the country's controversial 'social credit system', in which failure to comply with the rules of one government agency can mean facing restrictions or penalties from other agencies.

The punishment overhaul is the government's latest attempt to crack down on misconduct. But the nature and extent of the policy has surprised many researchers. "I have never seen such a comprehensive list of penalties for

research misconduct elsewhere in the world," says Chien Chou, a scientific-integrity education researcher at Chiao Tung University in Hsinchu, Taiwan.

Although some penalties for misconduct existed before the new policy — research programmes could be suspended and offenders could be barred from promotions — drawing them together under one framework makes them much more powerful, says Yang Wei, a former head of the National Science Foundation of China who is now an engineer at Zhejiang University in Hangzhou. Whether

"Almost all aspects of daily life for the guilty scientists could be affected."

the system will reduce misconduct will depend on how it is enforced, say some researchers. Others, including Chen, are certain it will work.

"Without doubt, it will be effective," he says.

The social credit system, which was introduced in 2014, has already had a large effect on life in the country. Failure to pay debts or fines can be recorded on the system's website and lead to restrictions when applying for a credit card, insurance or even train tickets.

As of April, the number of times people had been denied airline tickets as a result of the system had reached 11 million, and train tickets had been denied on 4.2 million occasions. More than 2 million people have paid debts or fines after facing these restrictions.

Chinese leaders have been increasingly focused on scientific misconduct, following

ongoing reports of researchers in the country using fraudulent data, falsifying CVs and faking peer reviews. In May, the government announced sweeping reforms to improve research integrity, including the creation of a national database of misconduct cases. Inclusion on the list could disqualify researchers from future funding or research positions, and might affect their ability to get jobs outside academia.

The punishment system seems to align with that goal. "It shows that China takes research integrity very seriously," says Max Lu, a chemical engineer and president of the University of Surrey in Guildford, UK, who has previously advised the Chinese government on science policy.

Lu thinks the system's success will depend on the resources that are devoted to enforcing it. The government is likely to focus on punishments for the most egregious cases first, such as repeat offenders, or those whose fraud has major consequences, says Li Tang, who studies science policy at Fudan University in Shanghai.

But the government needs to define what actions constitute major research misconduct, and how penalties will apply, says Chou.

Addressing misconduct in China will require more than punishments, says Tang. Mandatory courses on research integrity are becoming more common, but more could be done, she says. "Educating lab PIs and younger generations is extremely important," she says. ■

365 DAYS:
the year in science

Climate change is
intensifying droughts
in Australia.



2018

in review

A turbulent year marked by raging wildfires and allegations of bullying in the sciences comes to a close. But researchers can celebrate some milestones, including the most accurate map yet of the Milky Way's stars, and the discovery of bones showing that a woman who lived 90,000 years ago had a Neanderthal mother and a Denisovan dad.

UP IN SMOKE

Evidence of a changing climate continued to mount in 2018. More than 50 fires raged across Sweden in July, fuelled by intense heat and the driest conditions the country had seen in more than a century. By August, British Columbia in Canada was in the middle of its worst fire season on record, and California was battling the largest wildfire in its history. In November, the state faced its deadliest wildfire when the Camp Fire killed at least 85 people.

The situation will probably get worse. The Intergovernmental Panel on Climate Change released a report in October stating that, in as little as a decade, global temperatures could pass 1.5 °C of warming since pre-industrial times. And there is scant evidence that governments are taking aggressive action to combat global warming.

Australia's new prime minister, Scott Morrison, abandoned a policy in September that would have limited emissions from the electricity sector, a move that scientists said amounted to abandoning the nation's commitment to the 2015 Paris climate accord.

The US Environmental Protection Agency (EPA) proposed rolling back regulations intended to curb emissions from vehicles and power plants. And in April, then-EPA administrator Scott Pruitt released a proposal that would prevent the agency from basing regulatory decisions on data that aren't publicly available — potentially eliminating epidemiological studies that don't report health data owing to patient-privacy concerns.

On a more positive note, US President Donald Trump, who had gone longer without a top science adviser than any first-term president since at least 1976, finally nominated one in July. But the nominee, meteorologist Kelvin Droegemeier, was still awaiting Senate confirmation as *Nature* went to press. In China, the government created a ministry of ecological environment to track pollution and enforce environmental rules, as well as an agency to protect endangered species.

There was also movement on two groundbreaking lawsuits that seek to hold governments accountable for their inaction on climate change.

BROOK MITCHELL/GETTY

NSF

An appeals court in The Hague upheld a 2015 ruling in response to a lawsuit filed by environmentalists that holds the Dutch government responsible for cutting the nation's emissions to 25% below 1990 levels by 2020. And the US Supreme Court ruled in November that a case brought in 2015 by 21 young people against the US government can proceed. The plaintiffs argue that the government violated their rights to life, liberty and property by failing to prevent dangerous climate change.

POPULIST UPHEAVAL

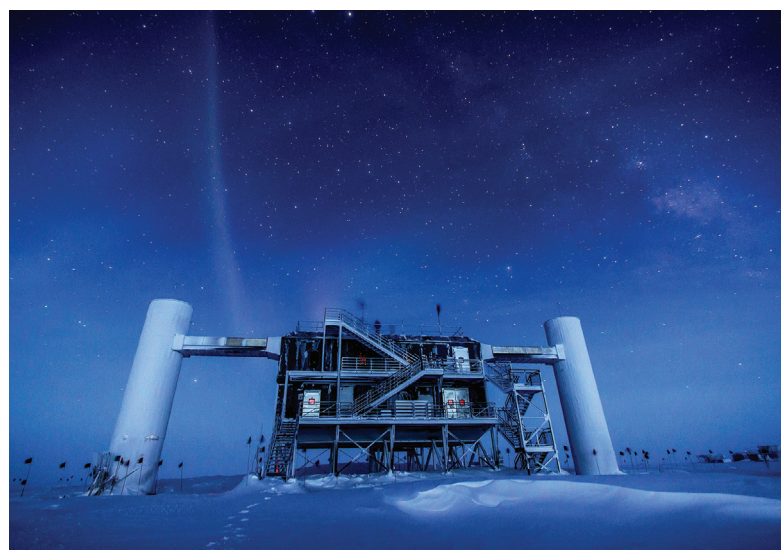
Brazilians elected right-wing candidate Jair Bolsonaro as their next president in October. He has promised to crack down on government corruption — but also to roll back environmental regulations. As a member of Brazil's lower house of Congress, Bolsonaro often voted with the conservative rural caucus, which sought to open up the Amazon rainforest to activities including farming. He takes office in January.

The political upheaval extended to Europe. In Italy, a coalition government comprised of two populist parties took over in June. The recently appointed health minister, physician Giulia Grillo, campaigned to roll back a 2017 decree making multiple vaccinations mandatory for schoolchildren. And in September, she announced that the government would probably continue to make only the measles vaccine compulsory.

In Hungary, Viktor Orbán's populist government announced that it would take control of the Hungarian Academy of Sciences budget, starting in 2019. What that means for scientists working in the academy's 44 research institutes is still unclear. Meanwhile, the Central European University (CEU) in Budapest, an international university founded by billionaire George Soros, was at the centre of a struggle between liberals and the government. An agreement that would allow the CEU to continue to fully operate in the country after the end of the year remains unsigned by the government. The university has announced that it will transfer most of its degree programmes to its Vienna campus in 2019, although research would continue at the Budapest campus.

And as *Nature* went to press, British Prime Minister Theresa May had just won a 'confidence vote' in her leadership, securing her position as head of the Conservative Party for another year. That vote was triggered by her decision to delay a crucial parliamentary vote on the unpopular Brexit deal hammered out by UK and European Union negotiators. May now hopes to discuss contentious parts of the deal with EU officials, who have insisted that the agreement itself can't be changed. The UK government plans to step up its preparations for the possibility of the country leaving the EU without any deal — a situation that could lead to UK scientists losing access to more than £1 billion (US\$1.3 billion) in annual EU research funding, and increased hurdles to the movement of staff, technologies and medicines in and out of the EU.

The political winds shifted in the United States after the country's



The IceCube observatory recorded data that could be used to track cosmic rays.

November midterm elections. The Democrats regained control of the House of Representatives — but not the Senate — from the Republican Party. The newly elected representatives, including at least 12 with backgrounds in science, technology, engineering or medicine, take office in January. The change will give Democrats control of key committees and the power to subpoena documents and testimony from President Donald Trump's administration. The incoming chair of the House's science panel, Representative Eddie Bernice Johnson (Democrat, Texas), has pledged to investigate the Trump administration, defend science from "political and ideological attacks", and address climate change.

THE RIGHT ANGLE

Back in the lab, a surprising property of graphene could help to solve a 30-year-old physics mystery. Two layers of the single-atom-thick form of carbon, when sandwiched together and offset by 1.1°, can mimic the superconducting behaviour of some copper-based materials called cuprates. The discovery gives physicists hope that they can use graphene to determine why cuprates conduct electricity without resistance at relatively warm temperatures (see page 325). Graphene is much better understood, and easier to manipulate, than cuprates. The finding, reported in March, could aid in the search for superconductors that don't need to be chilled close to absolute zero.

Superconductor researchers weren't the only physicists having a good year. In October, the European Commission unveiled the first round of winners in its 10-year, €1-billion (US\$1.1-billion) funding spree for quantum technologies. The 20 projects cover topics such as atomic clocks and secure communications. Meanwhile, the United Kingdom renewed its domestic quantum-hubs programme with an extra £235 million, and Germany pledged €650 million for quantum research over 4 years.

The end of the year saw the most significant overhaul of the standard units of measurement since 1875. In November, the General Conference on Weights and Measures in Versailles, France, approved a plan to define all units using fundamental constants of nature, rather than physical reference objects. For example, the kilogram is now rooted in the Planck constant of quantum mechanics rather than in 'Le Grand K', a platinum-alloy cylinder kept in a vault outside Paris.

CONTROVERSIAL EDITS

The year brought controversial news in the field of genetics. In November, Chinese scientist He Jiankui stunned the world by claiming the birth of the first gene-edited babies. His team used the CRISPR-Cas9 system to alter the *CCR5* gene, which encodes a protein that HIV uses to enter cells. The edited embryos produced twin girls, but it is unclear whether the changes will confer resistance ▶

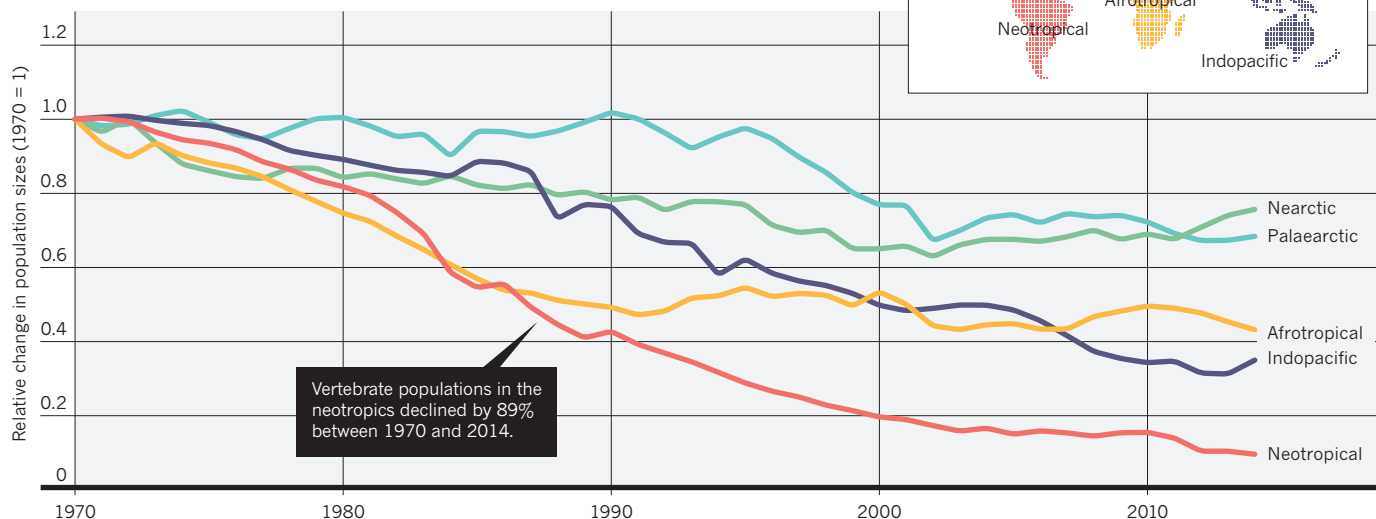


A man shows his support for Brazil's president-elect, Jair Bolsonaro.

CARL DE SOUZA/AFP/GETTY

DECLINING FORTUNES FOR ANIMAL POPULATIONS

A biennial analysis of land and freshwater animal species, released this year, finds that habitat destruction is one of the most common threats to their populations. Tropical regions were the hardest hit, based on average population sizes in 2014 — the most recent year for which data were available.



SOURCE: WWF/ZSL

► to HIV, especially because one of the babies seems to still have an intact copy of the gene.

Scientists the world over decried the work, warning that the technique is not ready for use in people. As *Nature* went to press, the Guangdong health commission was investigating He, and China's national science ministry had ordered him to stop doing research.

That announcement capped a year of genetic advances, including the first primates to be cloned using a method similar to the one used to produce Dolly the sheep. The breakthrough, announced in January, could eventually allow researchers to use gene editing to modify primate clones and create models of human disorders.

Another first centred on a young woman who lived some 90,000 years ago. She inherited half of her chromosomes from her Neanderthal mother, and the other half from a Denisovan dad, scientists revealed in August. Dubbed Denny, the hybrid woman is the only known first-generation offspring of two distinct hominin groups.

August also saw the approval of the first therapy that relies on a technique called RNA interference to silence a specific gene. It was the culmination of 20 years of dogged pursuit by researchers. The US Food and Drug Administration approved the drug to treat a rare disease called

hereditary transthyretin amyloidosis, which can lead to organ damage.

On the legal front, a fierce patent battle entered its end game in September, when a US federal appeals court upheld patents on CRISPR-Cas9 editing from the Broad Institute of MIT and Harvard in Cambridge, Massachusetts. The fight had pitted the Broad against another team of researchers from institutions including the University of California, Berkeley.

A July ruling by Europe's highest court placed gene-edited crops under the same strict regulations as conventional genetically modified crops: a potential setback for researchers who work on such organisms.

And a surprising turn in a cold case put genetic sleuthing in the spotlight in April. A public genealogy site called GEDmatch enabled an arrest in California's Golden State Killer case. Joseph James DeAngelo is accused of committing a string of murders, sexual assaults and robberies in the 1970s and 1980s. Investigators identified him in part by matching crime-scene DNA to genetic profiles posted by some of his distant relatives on GEDmatch.

BAD BEHAVIOUR

In the United Kingdom, researchers at several prominent institutes spoke out about bullying, while major science funders cracked down on workplace harassment. In May, the Wellcome Trust in London, a biomedical research funder, introduced a pioneering anti-bullying policy. Three months later, it revoked £3.5 million from cancer geneticist Nazneen Rahman, who had resigned from the Institute of Cancer Research in London following an investigation into bullying allegations. Rahman said at the time that she and her team would complete their Wellcome-funded research before she left the institute in October.

Later in August, complaints surfaced about the management of the Wellcome Sanger Institute in Hinxton, UK, which is funded by the Wellcome Trust. They included allegations that the institute's director, geneticist Michael Stratton, bullied staff, discriminated against them and misused funds. In October, an investigation reported failings in how people were managed, but cleared senior leaders, including Stratton, of wrongdoing. Stratton apologized for "failures in people management" and the resulting "unintended detrimental effects on individuals". The whistle-blower and some of those who made complaints disputed the investigation's findings.

Bullying was also an issue at the University of Bath, UK, which upheld a complaint against vertebrate palaeontologist Nicholas Longrich, who was part of the team that first discovered a fossilized four-legged snake. After the university's investigation, the Leverhulme Trust in



The suspect arrested in California's Golden State Killer case.

RANDY PENCH/SACRAMENTO BEE/TNS/GETTY

ESA/GAIA/DPAC

London revoked a nearly £1-million grant it had awarded to Longrich in 2016. He did not respond to *Nature's* requests for comment.

In the United States, sexual harassment continued to make headlines. A comprehensive analysis released by the US National Academies of Sciences, Engineering, and Medicine found that sexual harassment is pervasive in the country's academic sciences. The June report also concluded that a 1972 law prohibiting gender-based discrimination on US campuses that receive government funding, known as Title IX, had not reduced incidences of sexual harassment.

Accusations of harassment continued to rock several institutes. The president of the University of Rochester in New York resigned in January following campus protests over the university's handling of sexual-misconduct allegations against one of its professors, cognitive scientist Florian Jaeger. Jaeger has denied the accusations. Cancer researcher Inder Verma resigned from the Salk Institute for Biological Studies in La Jolla, California, in June after allegations of harassment, which he has denied. And three senior female faculty members who had sued the Salk over alleged gender discrimination — which they say harmed their careers — settled their cases with the institute.

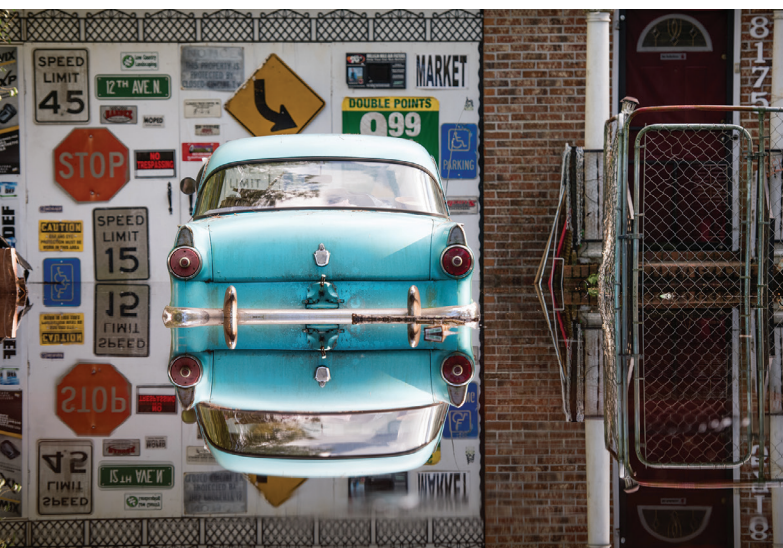
TEAR DOWN THAT PAYWALL

The launch of Plan S, a radical strategy to flip scholarly publications to a fully open-access model, rocked the world of publishing. Spearheaded by Robert-Jan Smits, the European Commission's open-access envoy, the programme started with the backing of 11 national research funders, including heavyweights from the United Kingdom, France and the Netherlands.

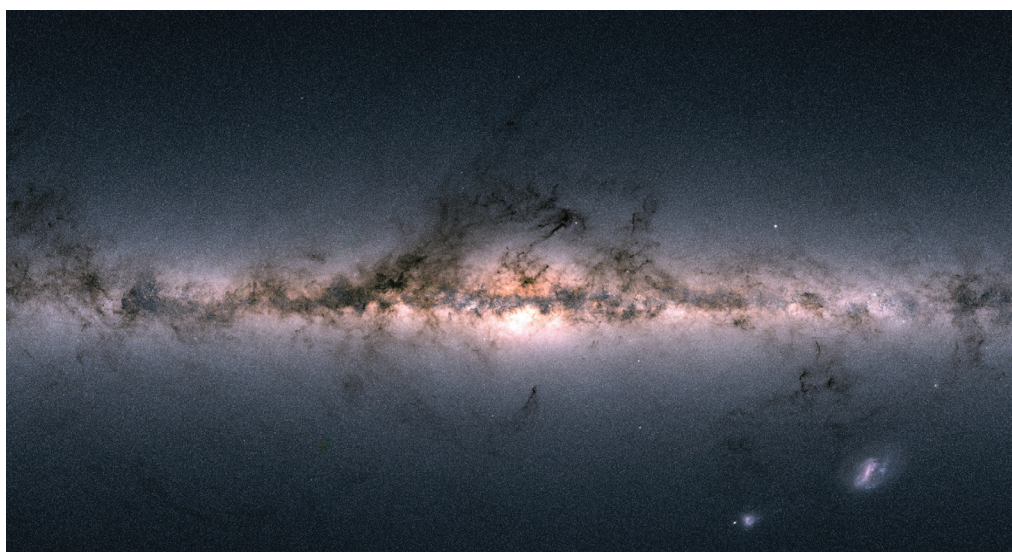
Beginning in 2020, the results of any research funded by Plan S backers — such as the Research Council of Norway and the Austrian Science Fund — must be freely available on publication, under a liberal publishing licence that allows others to reuse the work. Under certain conditions, the bold bid will allow work to appear in hybrid journals, which collect subscriptions while publishing some papers openly for a fee.

Finland joined Plan S shortly after the September launch, and major biomedical charities the Wellcome Trust, and the Bill & Melinda Gates Foundation in Seattle, Washington, signed up in November. China backed the plan in December.

SEAN RAYFORD/GETTY



Climate change could exacerbate flooding from storms.



Data from the Gaia mission helped to create a star map for the Milky Way.

FLY ME TO THE MOON

It was a year of beginnings and endings for the world's space agencies. NASA started developing concepts for a space station near the Moon this year, following a 2017 presidential order to return astronauts to the lunar surface. The agency is also working with companies to develop small lunar landers. And in December, China launched its Change-4 rover, which will attempt the first-ever soft landing on the Moon's far side.

The European Space Agency's (ESA's) BepiColombo mission launched in October on a journey to Mercury, and in August, NASA's Parker Solar Probe headed for the Sun. Meanwhile, two probes travelled into interplanetary space to gather cosmic dirt from near-Earth asteroids. The Japan Aerospace Exploration Agency's Hayabusa2 spacecraft dropped two small robots onto the asteroid Ryugu. And in December, NASA's OSIRIS-REx arrived at its own rock, named Bennu.

But the US space agency also said its share of farewells. Its Dawn spacecraft ran out of fuel in October after visiting the large asteroids Vesta and Ceres; in the same month, NASA ended science operations for its long-running exoplanet-hunter, the Kepler space telescope.

On Mars, a planet-wide dust storm in June cut off communications with NASA's 15-year-old Opportunity rover, which is now feared lost. But a discovery reported in July revealed a potential target for future exploration. Researchers announced that ESA's Mars Express orbiter had spotted a possible lake beneath the ice near the planet's south pole.

Back on Earth, two radio antennas in the Australian outback found indirect hints of the Universe's very first stars as they began to shine around 180 million years after the Big Bang. If scientists can confirm these signals of the 'cosmic dawn', announced in February, they'll have their first glimmers of an epoch that has so far been impossible to observe.

Data from ESA's Gaia probe yielded a 3D map of the Milky Way of unprecedented accuracy. It records the positions, distances, colours and speed and directions of motion of 1.3 billion stars, and has already led to more than 400 papers since its April release. The map has also demolished the image of the Milky Way as a steadily rotating spiral, showing instead that the Galaxy is still sloshing back and forth from interactions with smaller galaxies in the past one billion years.

And for the first time, astrophysicists traced the origins of a high-energy neutrino to a supermassive black hole at the centre of a distant galaxy. The finding, announced in July, could help researchers to pin down the source of cosmic rays — the most energetic particles in nature — because scientists think that some cosmic rays and high-energy neutrinos are produced in the same way. ■

Written by Alison Abbott, Ewen Callaway, Davide Castelvecchi, Holly Else, Elizabeth Gibney, Heidi Ledford, Jane J. Lee, Lauren Morello, Jeff Tollefson and Alexandra Witze.

IMAGES OF THE YEAR

2018 will go down in history as a scorcher: deadly wildfires and droughts raged from California to Cape Town. The year also brought advances in cloning and imaging — and a bleak reminder of the fragility of some of Earth's rarest species. Here are the striking shots from science and the natural world that caught the eyes of *Nature's* editors.

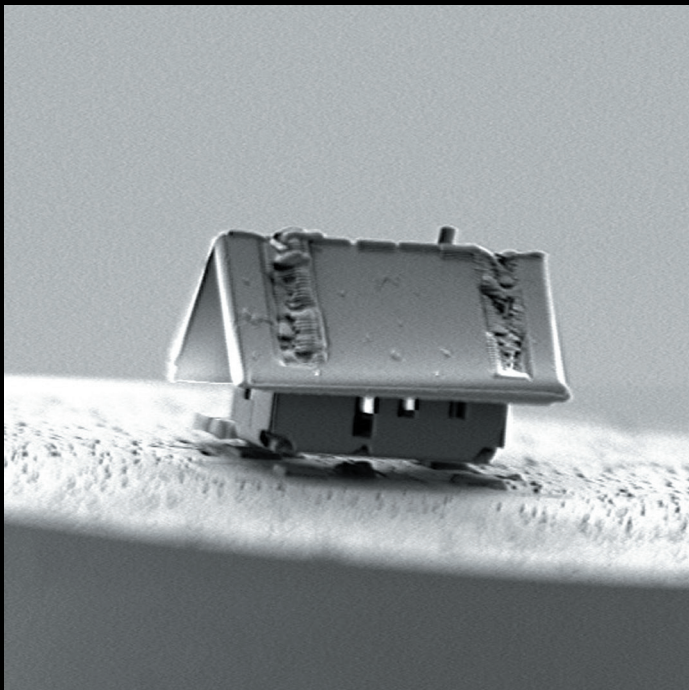
Images selected by *Nature's* art editors
Text by Mico Tatalovic



365 DAYS:
the year in science

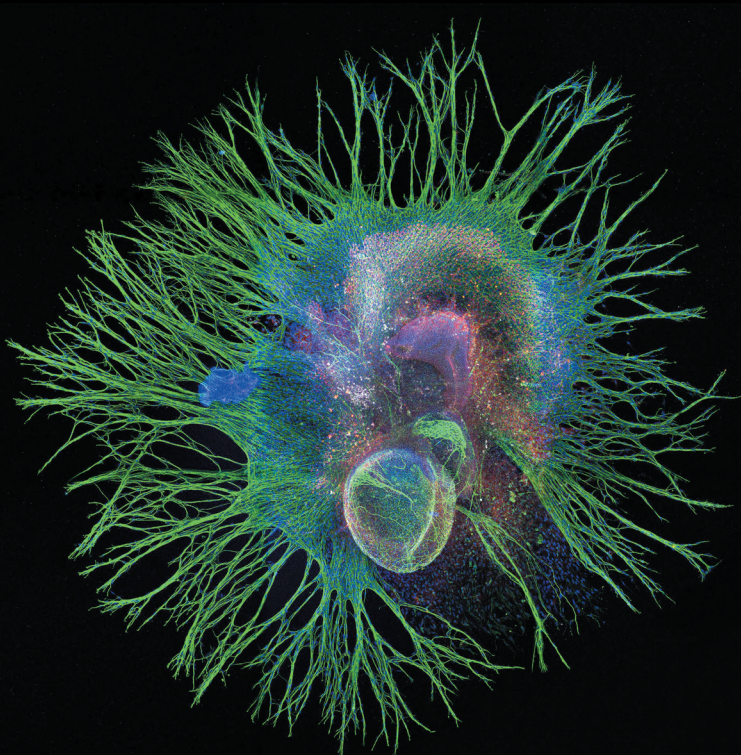
EXTREME DOWNSIZING

In May, a team at the Femto-ST Institute in Besançon, France, used nanoassembly tools — a focused ion beam, a gas-injection system and a tiny, manoeuvrable robot — to build this 20-micrometre-long house from silica.



SOUND SYSTEM

Cell biologists Stephen Freeman and Laurence Delacroix at Liège University in Belgium won distinction in the Nikon Small World Photomicrography Competition with this image of neurons in a mouse's inner ear. The neurons are cultured *in vitro* to study how neurons mature and become damaged.



STORM AND SWIRL

NASA's Juno spacecraft, now in the eighth year of its mission to Jupiter, offered rich data and spectacular images of the gas giant. Swirling clouds and a large storm — the white oval — are seen here in the planet's dynamic northern hemisphere.



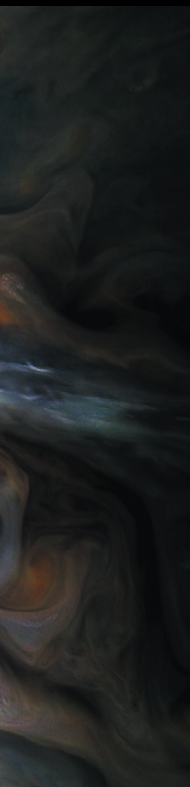
TO INFINITY

US company SpaceX continued its dominance in the commercial spaceflight arena, making a suite of rocket launches and landings. This February's launch from California carried a radar satellite and two Starlink satellites — part of the firm's ultimate goal to provide Internet worldwide.



X-RAY VISION

In September, ecologist W. Leo Smith at the University of Kansas in Lawrence, published a new imaging technique, used on this roosterfish (*Nematistius pectoralis*). The method involves stripping away an organism's muscles and staining its bones.



DYING DAYS

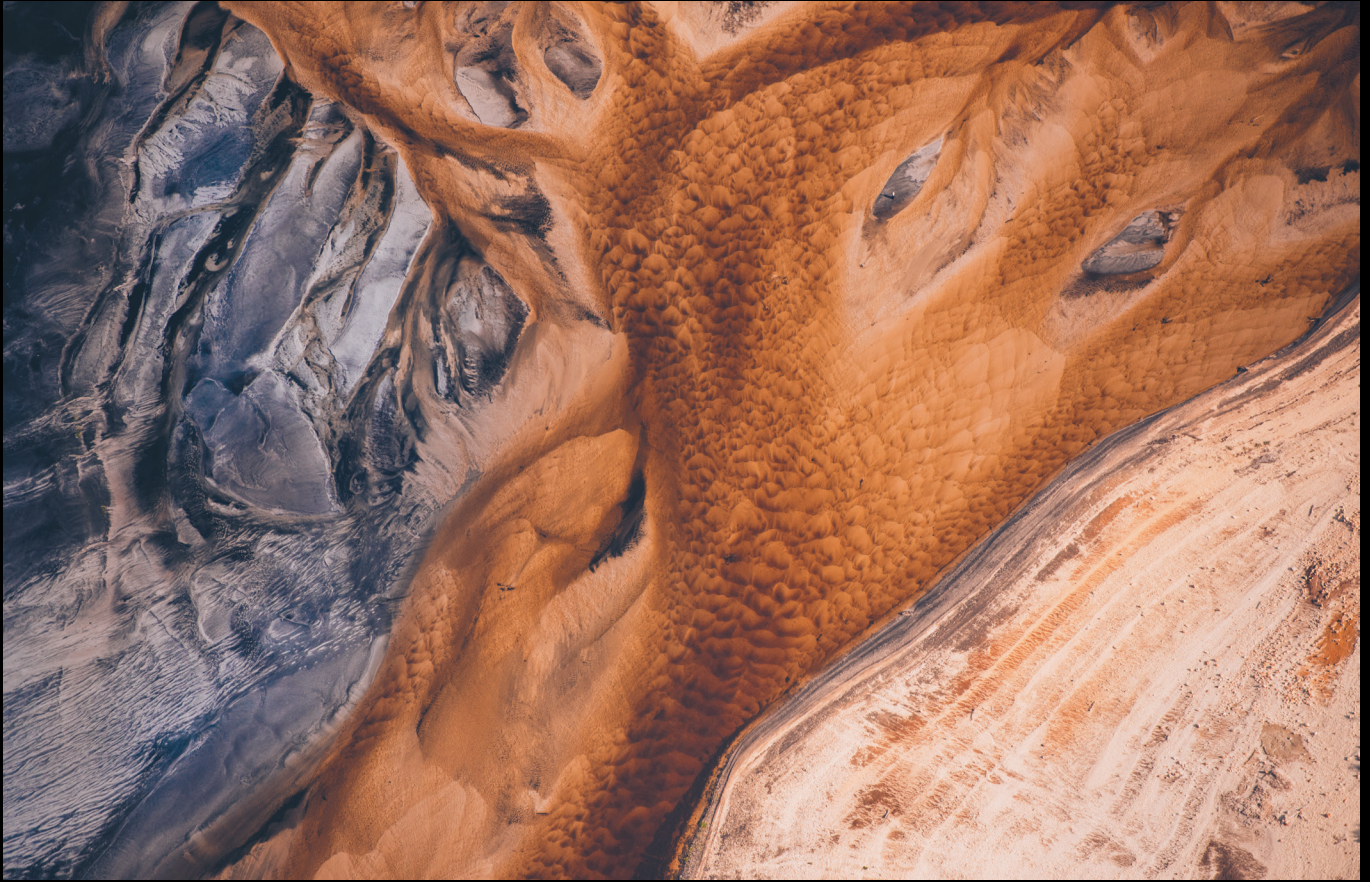
The world's last male northern white rhino, Sudan, died in March in Kenya. Only two females survive — and researchers are exploring ambitious *in vitro* fertilization techniques to save the subspecies (*Ceratotherium simum cottoni*).



THE SCORCHED STATE

California's wildfires — some of the state's largest on record — dominated headlines in 2018. Here, an aeroplane drops fire retardant in an area north of San Francisco in August. Authorities evacuated thousands of people.





SOUTH AFRICA'S CRIPPLING DROUGHT

Three years of record-breaking drought in South Africa prompted officials in Cape Town to consider a dramatic move: shutting off taps completely. City resident and photographer Kelvin Trautman captured the scale of the crisis in this image of an empty reservoir at Steenbras Upper Dam.



SMALL-WORLD SPORES

This 10x-magnified ultraviolet image of a fern sorus — the structure that produces and contains the plant's spores — won Rogelio Moreno Gill second place in the Nikon Small World Photomicrography Competition.

TWO OF A KIND

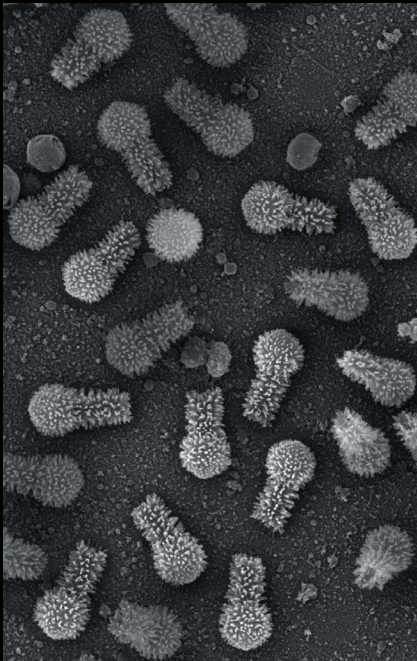
Meet Zhong Zhong and Hua Hua. The macaque twins, introduced to the world by Chinese researchers in January, were the first primates to be born using a cloning technique similar to that used to produce Dolly the sheep. Primates had proved difficult to copy using standard techniques.



WILDFIRE AIRDROP: MARK MCKENNA/ZUMA PRESS/PA. TINY HOUSE: FEMTO-ST/CATERS. SPACEX LAUNCH: SPACEX. X-RAY FISH: MATTHEW G. GIRARD. MOUSE NEURON: STEPHEN FREEMAN AND LAURENCE DELACROIX/NIKON SMALL WORLD 2018. JUPITER STORM: NASA/JPL-CALTECH/SWRI/MSSS/GERALD EICHSTADT/SEAN DORAN. WHITE RHINO: AMI VITALE/NATIONAL GEOGRAPHIC CREATIVE. SA DROUGHT: KELVIN TRAUTMAN. SMALL-WORLD SPORES: ROGELIO MORENO GILL/NIKON SMALL WORLD 2018. CLONED MONKEYS: JIN LIWANG/XINHUA VIA ZUMA. GIANT VIRUS: J. ABRÁHÃO ET AL/NATURE COMMUN. ICE FOOTBALL: MARIUS VAGENES VILLANGER/NTB SCANPIX VIA REUTERS. DIVING BIRDS: GREG LECOEUR/UPY 2018.

GIANT GENOME

A newly discovered giant Tupanvirus, found in amoebae, has both the longest tail and the largest set of genes involved in protein-making of any known virus.



WARMING UP

An Arctic game of football is a risky business. During a match in March on an ice floe near Greenland, armed guards watched for polar bears while scientists from Norway's Institute of Marine Research and crew from a naval ice-breaker played.

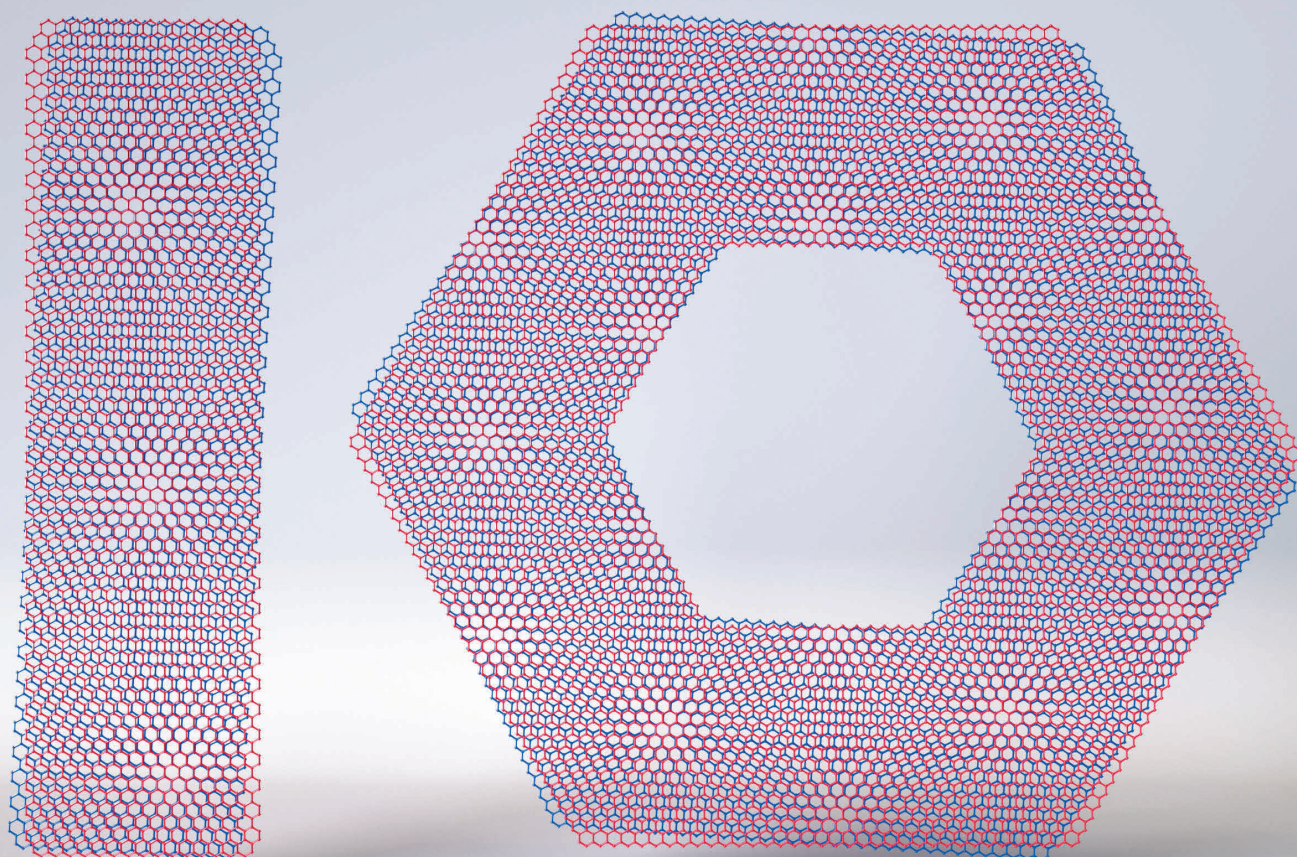


DEEP DIVE

Gannets in Scottish waters dive to hunt for mackerel and other fish. The birds drop from a height of 30 metres, achieving speeds of 100 kilometres per hour. The image won third place in the behaviour category of the Underwater Photographer of the Year Competition.

NATURE'S 10

Ten people who mattered this year



YUAN CAO / VIVIANE SLON / HE JIANKUI / JESS WADE
VALÉRIE MASSON-DELMOTTE / ANTHONY BROWN / BEE YIN YEO
BARBARA RAE-VENTER / ROBERT-JAN SMITS / MAKOTO YOSHIKAWA

About the image:

This design highlights advances in studies of atom-thick materials with unusual properties. The image represents two graphene sheets offset by a 'magic' angle, an arrangement that can behave as a superconductor in certain conditions. Image by JVG.

365 DAYS:
the year in science

YUAN
CAO

GRAPHENE WRANGLER

A PhD student coaxed superconductivity from sheets of atom-thick carbon.

BY ELIZABETH GIBNEY

Yuan Cao's teenage years were hardly typical. By age 18, he had already graduated from high school, completed an undergraduate degree at the University of Science and Technology of China in Hefei, and travelled to the United States to begin his PhD. He hasn't slowed down since: this year, aged just 21, Cao had two papers published on strange behaviour in atom-thick layers of carbon that have spurred a new field of physics. Cao admits that his situation is unusual, but says he isn't special. After all, he did spend a full four years at university: "I just skipped some of the boring stuff in middle school."

Pablo Jarillo-Herrero's group at the Massachusetts Institute of Technology (MIT) in Cambridge was already layering and rotating sheets of carbon at different angles when Cao joined the lab in 2014. Cao's job was to investigate what happened in two-layer stacks when one graphene sheet was twisted only slightly with respect to the other, which one theory predicted would radically change the material's behaviour.

Many physicists were sceptical about the idea. But when Cao set out to

create the subtly twisted stacks, he spotted something strange. Exposed to a small electric field and cooled to 1.7 degrees above absolute zero, the graphene — which ordinarily conducts electricity — became an insulator (Y. Cao *et al. Nature* **556**, 80–84; 2018). That by itself was surprising. "We knew already that it would have a big impact on the community," says Cao. But the best was yet to come: with a slight tweak to the field, the twisted sheets became a superconductor, in which electricity flowed without resistance (Y. Cao *et al. Nature* **556**, 43–50; 2018). Seeing the effect in a second sample convinced the team that it was real.

The ability to coax atom-thick carbon into a complex electronic state through a simple rotation now has physicists clamouring to engineer exciting behaviour in other twisted 2D materials. Some even hope that graphene could shed light on how more-complex materials superconduct at much higher temperatures. "There are so many things we can do," says Cory Dean, a physicist at Columbia University in New York City. "The opportunities at hand now are almost overwhelming."

Hitting graphene's 'magic angle' — a rotation between parallel sheets of around 1.1° — involved some trial and error, but Cao was soon able to do it reliably. His experimental skill was crucial, says Jarillo-Herrero. Cao pioneered a method of tearing a single sheet of graphene so that he could create a stack composed of two layers with identical orientation, from which he could then fine-tune alignment. He also tweaked the cryogenic system to reach a temperature that allowed superconductivity to emerge more clearly.

Cao loves to take things apart and rebuild them. At heart, he is "a tinkerer", his supervisor says. On his own time, this means photographing the night sky using homemade cameras and telescopes — pieces of which usually lie strewn across Cao's office. "Every time I go in, it's a huge mess, with computers taken apart and pieces of telescope all over his desk," says Jarillo-Herrero.

Despite his youth and shy manner, colleagues say that Cao's maturity shines through in his persistence. Having missed out by a whisker on a

place in MIT's physics graduate programme, for example, Cao found a way to pursue the subject by joining Jarillo-Herrero's team through the electrical-engineering department. Cao also shrugged off a disappointing start to his PhD, after realizing that seemingly exciting data that he had spent six months trying to understand were due to a quirk of the experimental set-up. "He wasn't happy, but he just rolled up his sleeves and continued working," Jarillo-Herrero says.

Cao, now 22, doesn't yet know where he'd like his career to lead. "On magic-angle graphene, we still have a lot of things to do," he says. But universities around the world are already eyeing him for not only postdoctorate jobs, but also faculty positions, says physicist Changgan Zeng, Cao's undergraduate supervisor and mentor at the University of Science and Technology of China. "Among condensed-matter physicists in China, everybody knows his name," Zeng says. The university would gladly have him back, but Zeng expects that Cao will stay in the United States for now. "There, it's easier to see the stars." ■

HUMANITY'S HISTORIAN

A palaeogeneticist discovered a remarkable ancient hybrid hominin: half Neanderthal, half Denisovan.

BY EWEN CALLAWAY

Viviane Slon was sure she had made a mistake three years ago, when DNA tests on an ancient bone fragment pointed to a union of two extinct human groups. Half of the genome looked like a Neanderthal's; the other half matched sequences from Denisovans — a group once found throughout Asia.

“I was very much of the mindset that this cannot be,” says Slon, a palaeogeneticist at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany. Slon told no one for several days, and wondered whether she had made some mistake.

When she couldn't find an error, Slon shared the results with her colleagues and began to ponder what they might mean. Further tests determined that the individual — a young adult female affectionately named Denny by colleagues — was the daughter of a female Neanderthal and a male Denisovan who lived roughly 90,000 years ago.

Neanderthal and Denisovan genomes point to past interbreeding, but a direct product of such an encounter had never been found.

The discovery, reported in August, reverberated with other scientists and the public, triggering hundreds of news articles and thousands of tweets. “It's probably the most fascinating person who's ever had their

genome sequenced,” one geneticist said at the time.

“It was fun to see people who are not at all in this field and do not think about Neanderthals in their everyday life, how this caught their attention,” says Slon, a postdoctoral fellow working with palaeo-geneticist Svante Pääbo.

Slon's perspective is unique among her peers, says Israel HersHKovitz, a palaeoanthropologist at the University of Tel Aviv, Israel. He supervised some of Slon's graduate-degree research, which spanned archaeology, anthropology, pathology and anatomy; she even supported herself by working in a cadaver lab. “She was not born in a sterile DNA lab,” says HersHKovitz. “When she speaks about the Neanderthal, she sees the Neanderthal. She sees its physiology, its anatomy, not just its genes.”

Slon says she is drawn to using genetics and other scientific approaches to study prehistory because of the lack of written records. “Everything you can infer is from what people left behind,” she says. “It's almost like solving a riddle.”

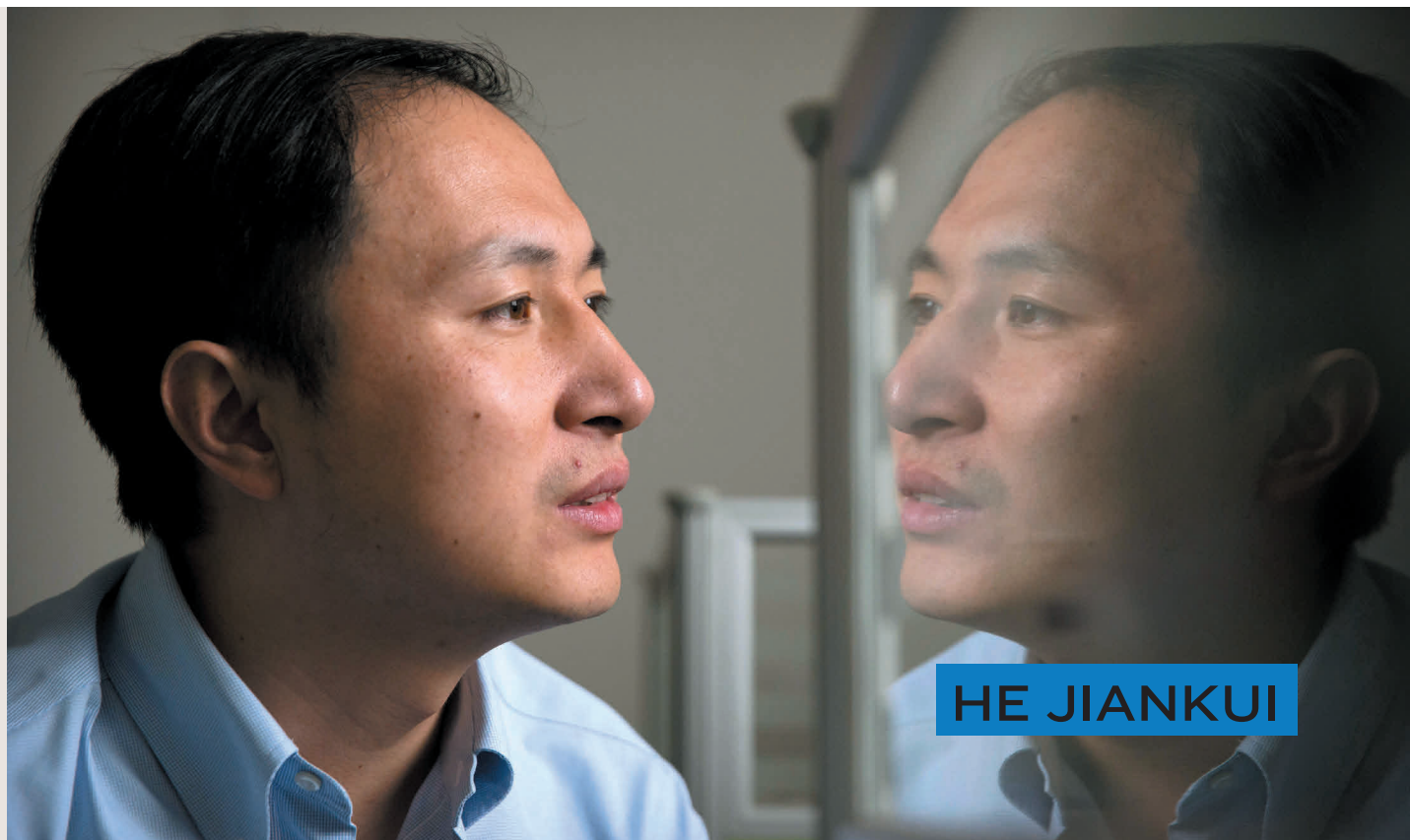
Much of her palaeogenetics research has centred around material from Denisova Cave, the vast cavern in southern Siberia that gave its name to the Denisovans, a cousin group to Neanderthals.

Slon's first project on remains from the cave was to sequence the DNA in a tooth from the fourth Denisovan individual found there. She also co-led a team that found Denisovan DNA in excavated dirt, an approach that could transform palaeogenetics — because it doesn't rely on finding rare hominin bones. Slon's colleagues had to screen more than 2,300 unidentified bone fragments to find Denny.

Slon is still working on material from Denisova Cave, which she got to visit for the first time earlier this year. And she will continue to work on extracting hominin DNA from sediments. She doesn't expect to happen on another once-in-a-lifetime find like Denny, but she is eager to plumb ancient genomes for all sorts of personal insights, such as family relationships between ancient humans or how living conditions influenced the individuals' health. She also hopes to examine the lives of hominins who lived beyond Eurasian sites. “There's a whole world that can still be explored,” she says. ■



VIVIANE SLON



HE JIANKUI

MARK SCHIEFELBEIN/AP/REX/SHUTTERSTOCK

CRISPR ROGUE

A scientist's claim to have created gene-edited babies generated international furore.

BY DAVID CYRANOSKI

He Jiankui knew he was crossing a new bioethical boundary, when he revealed in November that he had altered the genomes of two infants — in a way that would be passed on to future generations. “I understand my work will be controversial, but I believe families need this technology and I'm willing to take the criticism for them,” he said in a video announcing the births of twin girls whose genomes he had edited using CRISPR, ostensibly to protect them from HIV infection.

The reaction to the news was stronger than He had expected. He was widely criticized for ignoring important ethical considerations and exposing the girls to unknown risks for an uncertain benefit. The Southern University of Science and Technology in Shenzhen, China, where He works, distanced itself. The Chinese science ministry forbade him from continuing research. And the health ministry launched an investigation. He, who is now not speaking to the press,

disappeared from the world stage as quickly as he had emerged.

He came to gene editing as an outsider. The first publication listed on his website, from a decade ago, is related to quantum physics. In 2010, he had publications on economics, evolution and the nature of curious repeated sections of DNA in bacterial genomes. He won some acclaim for his work in genome sequencing. A company he founded, Direct Genomics in Shenzhen, targeted the clinical-sequencing market and pulled in hundreds of millions of dollars in investments.

But He wanted to get into gene editing. He visited Feng Zhang, a CRISPR pioneer, at his laboratory at MIT, who warned him against editing human embryos for reproduction. Mark DeWitt, a geneticist at the University of California, Berkeley, says that he advised the same. Jennifer Doudna at Berkeley, another CRISPR pioneer, refused He's request for a visit because she thought he wasn't doing anything related to this technology. Now, she wonders whether He was “trying to leave a trail” of reputable contacts so he could say that he had broad support.

He will leave a complicated legacy. Scientists worry that the field of gene editing might now struggle to secure funding, regulatory approval or support from the public. And although the technology could lead to new insights into human development and potentially some ways of preventing deadly genetic disorders, few would argue that He's approach has helped. “I think he will be judged harshly,” says DeWitt. ■

Additional reporting by Ewen Callaway



JESS WADE

DIVERSITY CHAMPION

A physicist wrote hundreds of Wikipedia pages to boost the profiles of scientists from under-represented groups.

BY NISHA GAINO

When Jess Wade started writing a Wikipedia page every day, she didn't expect her efforts to earn her global attention. She was simply trying to correct the online encyclopaedia's under-representation of women and people of colour in science. But in July, when she tweeted about a trollish comment she'd received about the work, it prompted an outpouring of support and a big boost for her quest. "That wouldn't have happened without that one mean comment," she says.

Wade, a polymer physicist at Imperial College London, has tackled many science-outreach projects aimed at fostering diversity. She took up her page-a-day habit after learning that 90% of Wikipedia editors are men and only about 18% of people profiled on the site are women.

She has now created about 400 pages and works with organizations to host regular 'edit-a-thons' — in which people create and edit Wikipedia content with an eye to inclusivity. These have inspired similar events around the world, including some focused on other professions.

The visibility and momentum that Wade's project created is important, says Lenna Cumberbatch, who studies diversity at the University of St Andrews, UK. Although Wikipedia entries won't fix science's inclusivity problems, efforts such as Wade's help to change people's expectations. "She's redressing an imbalance that's existed for aeons," says Cumberbatch. "When you're literally writing history — that's kind of cool."

Wade's Wikipedia campaign isn't the only thing that thrust her into the spotlight this year. In September, she spoke about her engagement work at a conference on gender at CERN, Europe's particle-physics lab near Geneva, Switzerland. On the same day, physicist Alessandro Strumia from the University of Pisa in Italy delivered a presentation questioning women's ability in physics and attacking policies that encourage diversity. "His presentation was totally inappropriate," Wade says, "telling a room of mainly young woman scientists that they'd only ever achieve success in physics due to affirmative action".

Wade once again used social media to highlight the comments, and they were widely condemned. Strumia has been suspended from his work with CERN while an investigation is ongoing.

Wade expects to press on with her outreach, including stocking school libraries with the book *Inferior* by Angela Saini, which explores the harm caused by gender stereotypes. "I think diverse teams do better science," she says. "Doing all this stuff definitely makes sure that the academic community is more robust, resilient and creative." ■

EARTH MONITOR

A climatologist was a driving force behind the IPCC's stark report on global warming.

BY JEFF TOLLEFSON

In October, Valérie Masson-Delmotte and her colleagues presented the world with alarming news about its future. Within as little as a dozen years, Earth's average temperature could reach 1.5°C above what it was in the mid-nineteenth century, triggering a wave of changes that would transform ecosystems and kill off most of the world's coral reefs, among many other impacts.

The warning came courtesy of a special report from the Intergovernmental Panel on Climate Change (IPCC), in which Masson-Delmotte played a primary part. A climatologist at the Laboratory for Sciences of Climate and Environment in Gif-sur-Yvette, France, and co-chair of the IPCC working group that assesses the physical science of climate change, Masson-Delmotte helped to gather the report's authors, coordinate their work and, ultimately, get the report approved by governments.

The IPCC normally takes the better part of a decade to produce its massive assessments, but the 1.5°C report came together quickly, incorporating research published just weeks before the final draft was submitted for government review. "I'm really proud," Masson-Delmotte says. "We had a horribly stringent timeline, but I think we managed to build trust and ownership of the report by the authors."

The report makes clear that limiting warming to 1.5°C would have huge benefits compared with allowing temperatures to surge

to the 2°C level. But keeping to 1.5°C would require aggressive action to curb greenhouse-gas emissions. And even if nations could somehow achieve that, the world would look very different: entire ecosystems could be destroyed across more than 6% of the planet's terrestrial surface, and 70–90% of coral reefs would probably disappear.

"This report will be a hard one to ignore," says co-author Ove Hoegh-Guldberg, who is director of the Global Change Institute at the University of Queensland in St Lucia, Australia.

Diana Liverman, a geographer at the University of Arizona in Tucson, singles out Masson-Delmotte's work to improve diversity and representation in the IPCC. Women made up just 22% of the author team on the last assessment, completed in 2014; in this report, they comprised an unprecedented 40%. Masson-Delmotte also worked to elevate the role of early-career scientists and researchers from the global south. And for the next full climate assessment, due out in 2021, she has introduced procedures to promote engagement by all authors — including an online participation tool for scientists who are uncomfortable speaking up during meetings.

In an attempt to break down scientific silos, researchers from various disciplines worked together on every chapter. The result, Masson-Delmotte says, was an analysis that focused less on emissions scenarios and more on social, technological and governmental policies that could foster change — without exacerbating poverty and inequality around the world.

Masson-Delmotte spent ten days talking about the report and the wider IPCC process with delegates at the United Nations climate summit in late 2018. Now, she and the other co-chairs are pushing forward with two more reports — one on terrestrial biomes, the other on oceans and polar regions, slated for release in August and September 2019, respectively.

LAURENCE GEAR FOR NATURE

Similar to the IPCC itself — participation in which is a voluntary affair — Masson-Delmotte says that she is stretched to the limit. Her own research has been relegated to occasional nights, weekends and train rides, and she doesn't see as much of her two daughters and husband as she used to. "It's frustrating," she says. "But at the same time, it's awfully stimulating." ■



ANTHONY BROWN

STAR MAPPER

Working behind the scenes, an astronomer coordinated the release of Gaia's long-awaited bounty of Milky Way data.

BY RACHEL COURTLAND

For many astronomers, Christmas this year came on 25 April at precisely 10:00 Coordinated Universal Time. That was when scientists with the European Space Agency's Gaia mission published its first major data set: a 551-gigabyte catalogue detailing the positions and movements of more than 1.3 billion stars.

Researchers around the world were eager to dive into the data. But Anthony Brown, an astronomer at Leiden Observatory in the Netherlands, had a different feeling when the catalogue finally rolled out: "Tired," he says.

Brown had good reason. He leads the Gaia project's Data Processing and Analysis Consortium, a group of more than 400 researchers that had been crunching the numbers for years. The Gaia spacecraft, which launched in 2013, spins to scan the sky and records the starlight that streaks across the camera. Boiling the craft's data down into precise information on stellar positions, motion and other properties requires sophisticated processing on the ground.



VALÉRIE MASSON-DELMOTTE

To researchers who are more interested in using Gaia to explore the mysteries of the Milky Way, Brown's job might seem less than glamorous. A calm and measured character, Brown has worked as the data-processing consortium's chair since 2012. His day-to-day job is intensively administrative: much of his time involves coordinating with and meeting consortium teams to make sure that the mission's data-crunching pipeline, which fans out from an operations centre near Madrid, works smoothly.

But Brown's care and expertise have been crucial to the success of Gaia's data set, which has already been cited in more than 700 research papers. His efforts have helped to steer the collaboration through myriad snags, including a systematic error in the telescope's parallax data — measures of angles to stars that enable astronomers to work out distances. The team decided to characterize the problem carefully and explain it in the release, rather than delay for more than a year to collect more data to reduce the error, says Amina Helmi, an astronomer at the Kapteyn Astronomical Institute in Groningen, the Netherlands, and a member of the consortium. Brown has an impressive ability to motivate researchers who would rather be working on science, Helmi says. "I don't know how he does it," she says. "We all respect him. There is really this feeling that we all share that wants to make Gaia a success."

Brown and his colleagues have had little time to catch their breath. They are already preparing the next data release, which will probably be ready some time in the first half of 2021. Another is planned after that, and more could follow: in November, the mission was extended to at least the end of 2020. Brown, who has been involved with Gaia since 1997, is in no rush to see it end: "Having worked on this mission more than 20 years now, it's definitely part of who I am." ■

TIMOTHY ARCHIBALD FOR NATURE

BEE YIN
YEO



FORCE FOR THE ENVIRONMENT

Malaysia’s new science and environment minister became a strong voice against plastic pollution.

BY YAO-HUA LAW

Bee Yin Yeo began to question the future of the world — and her own career — while evaluating oil wells in the deserts of Turkmenistan. The new university graduate decided that humanity would eventually move away from fossil fuels, so she decided to find another profession that would serve the well-being of the world.

A few years later, in 2010, she returned home to Malaysia, armed with a master’s degree in advanced chemical engineering from the University of Cambridge, UK. She joined politics and won a seat in a state legislative assembly in 2013. Then, a political tsunami hit Malaysia: on 9 May 2018, voters ousted the coalition that had held uninterrupted power since the country’s founding in 1963. The new government brought in its own cabinet members, including Yeo, who was appointed Minister of Energy, Science, Technology, Environment and Climate Change.

Yeo was “shocked” when she first heard of her appointment. “It was unimaginable,” said the 35-year-old, who grew up in a small town amid oil-palm and rubber-tree estates in southern Malaysia. Yeo had spent

the previous 5 years attacking national policies, and now she could change them.

Since taking office on 2 July, Yeo has made several bold steps in reforming how Malaysia manages its environment and research. She announced goals to increase renewable energy from 2% to 20% of total energy generation by 2030, to reform the electricity market and to ramp up energy efficiency. She also went to battle against plastics pollution — which plagues southeast Asia. She criticized the influx of plastic waste into Malaysia, and helped to set a nationwide ban on its import. On 31 October, Yeo launched a 12-year roadmap and legal framework towards eliminating single-use plastic in Malaysia by 2030, which also calls for research and commercialization of eco-friendly alternatives, such as biodegradable plastics.

Yeo’s efforts parallel an escalating global concern over single-use plastics. In October, the European parliament voted to ban their use in products such as straws and cutlery. And a growing number of other nations have issued similar bans.

Julian Hyde, general manager of the environmental organization Reef Check Malaysia in Kuala Lumpur, praises Yeo’s efforts and roadmap. “The most important thing about it is that it’s over a realistic timescale.”

But the Malaysian Plastics Manufacturers Association (MPMA) sees problems ahead. Ching Yun Wee, who chairs the MPMA’s sustainability subcommittee, says that local manufacturers can now produce biodegradable plastics, but that the material cannot yet decompose as quickly or completely as is needed to solve the problem of plastic pollution.

Wee says, however, that compared to her predecessors, Yeo has given the MPMA more opportunities to voice its opinion.

Yeo says that by funding local research and adopting foreign techniques, Malaysia can develop the technology for biodegradable plastic. “Some people think of problems to solutions, and not solutions to the problem,” she says. “When business as usual is not possible, you find another solution.” ■

VINCENT PAUL YONG

BRIAN L. FRANK/NT/REDUX/EVINE



BARBARA
RAE-VENTER

DNA DETECTIVE

A genealogist helped to identify a serial killer and paved the way for DNA to play a larger part in solving crimes.

BY BRENDAN MAHER

In February 2017, Barbara Rae-Venter got a call from an investigator looking for help with a criminal case. “I said, ‘Sure,’” says Rae-Venter, a retired patent attorney in northern California, unaware that she was signing up to try and catch one of the most notorious serial killers and rapists in US history. This year, Rae-Venter’s work not only led to the killer’s arrest, but also demonstrated a powerful — if controversial — approach for identifying criminals through genetic genealogy.

“She opened the door for others who wanted to do this, but had reservations,” says CeCe Moore, who heads a forensic-genealogy unit at the company Parabon Nanolabs in Reston, Virginia.

Rae-Venter first trained in genetic genealogy — which uses DNA to fill out family trees — to explore her own ancestry. Eventually, she started using the tools to aid others, such as people who had been adopted as children, which drew the attention of Paul Holes, an investigator with the Contra Costa county district attorney’s office in California.

Holes was on the trail of a man who had terrorized California during the 1970s and 1980s. With 12 murders, 45 rapes and 120 burglaries attributed to him, the elusive perpetrator had become known as the East Area Rapist, the Original Night Stalker and the Golden State Killer.

Holes reasoned that if Rae-Venter could piece together the killer’s family history, it could help to find his true name.

Rae-Venter uploaded a profile made from crime-scene DNA into GEDmatch, a public database used by genealogists. Although not nearly as large as commercial genealogy websites, GEDmatch’s terms of service didn’t expressly prohibit law enforcement from doing searches.

Right away, she found someone who seemed to be a third or fourth cousin to the killer. With the help of the FBI and local law officials, she worked to triangulate a common ancestor and then build the family tree. She eventually zeroed in on Joseph DeAngelo, a former police officer living in Sacramento. A direct test of his DNA proved the match.

Many in the genealogy community knew that this approach was possible and there had been ongoing debates over whether it constituted an invasion of privacy. Moore says that she had been approached in the past to help in this way, but declined because of the debate and because most people who used GEDmatch were unaware that it could be done. DeAngelo’s highly publicized arrest changed that: the genealogy community, by and large, embraced this use of data, at least for finding violent criminals.

Curtis Rogers, a co-founder of GEDmatch, has amended the database’s rules to make it clearer that law enforcement might use the information. He hasn’t seen a mass exodus from his site, he says.

The floodgates have now opened for these kinds of cases. Under Moore’s direction, Parabon Nanolabs has uploaded about 200 perpetrator profiles to GEDmatch, resulting in at least 22 identifications and nearly as many arrests.

Rae-Venter says that she has been approached for help in more than 70 cases. Quiet and private, she is nevertheless excited to get more involved. After all, her new calling seems to run in the family. In her own research, she identified a great uncle who was a detective inspector with the London Metropolitan Police during the time of Jack the Ripper. “I would love to find out which cases he worked on,” she says. ■

ONES TO WATCH

2019

JEAN-JACQUES MUYEMBE-TAMFUM

Director-general of the Democratic Republic of the Congo National Institute for Biomedical Research

As his nation battles a worsening Ebola outbreak, this veteran virologist is spearheading the deployment of experimental therapies and a new vaccine.

JULIA OLSON

Co-counsel for the plaintiffs in *Juliana v. United States*

This lawyer is suing the US government on behalf of people who claim that the country has violated their rights by not preventing climate change.

MUTHAYYA VANITHA

Project director of India's Chandrayaan-2 Moon mission

A big moment for this engineer could come in early 2019, as India plans to land a rover near the lunar south pole and explore that region for the first time.

MAURA MCLAUGHLIN

Chair of management team at the North American Nanohertz Observatory for Gravitational Waves

This astronomer and her colleagues monitor neutron stars, and could soon detect gravitational waves created by supermassive black holes for the first time.

SANDRA DÍAZ

Co-leader of the Global Assessment of Biodiversity and Ecosystem Services

Díaz and researchers from more than 50 countries will release a major biodiversity report as part of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services.

OPEN-ACCESS LEADER

A bureaucrat launched a drive to transform science publishing.

BY HOLLY ELSE

The architect of this year's bold push to get rid of paywalls in science publishing says he got his ideas from an unlikely source: the publishers themselves.

In March, Robert-Jan Smits was tasked by the European Union's research commissioner, Carlos Moedas, with a special one-year mission: to get more research papers published outside journal paywalls, and fast. A veteran science-policy bureaucrat, Smits decided to go to the source: he asked big publishers how he could do it. They told him that if the organizations that pay for research insisted the findings had to be published openly, journals would have to adapt.

So that's what Smits set out to persuade research funders to do — in a plan launched in September that has sent shock waves through science publishing.

Smits has spent decades pulling the science-policy strings at the European Commission, and, until his current assignment, had served eight years as the director-general of research. He was ideally connected to begin rallying Europe's agencies with the idea, dubbed Plan S for 'science, speed, solution, shock', as he puts it. As *Nature* went to press, 16 funders had signed the plan; they require that the results of work they support be made freely available at the time of publication, starting in 2020.

Publishers have been dictating how research is published for decades, Smits says. "Now it is the funders calling the shots, and we will do things differently."

It's too early to know what the ultimate impact of Plan S on research publishing will be. Its details are open for consultation, and much might depend on how many other funders adopt the idea — but it will at least improve access to research, says Peter Suber, director of the Harvard Open Access Project and the Harvard Office for Scholarly Communication in Cambridge, Massachusetts. Smits has been overwhelmed with messages of support. But the initiative has also met with resistance: several publishers have said it could put them out of business, and some researchers have said that they don't want their choice of where to publish to be restricted.

Smits is no stranger to disrupting the status quo in European science. In 2007, he was instrumental in setting up the excellence-focused European Research Council (ERC) funding agency — when, he says, very few member states wanted it. "We had to go country by country to convince people that we needed it," he says.

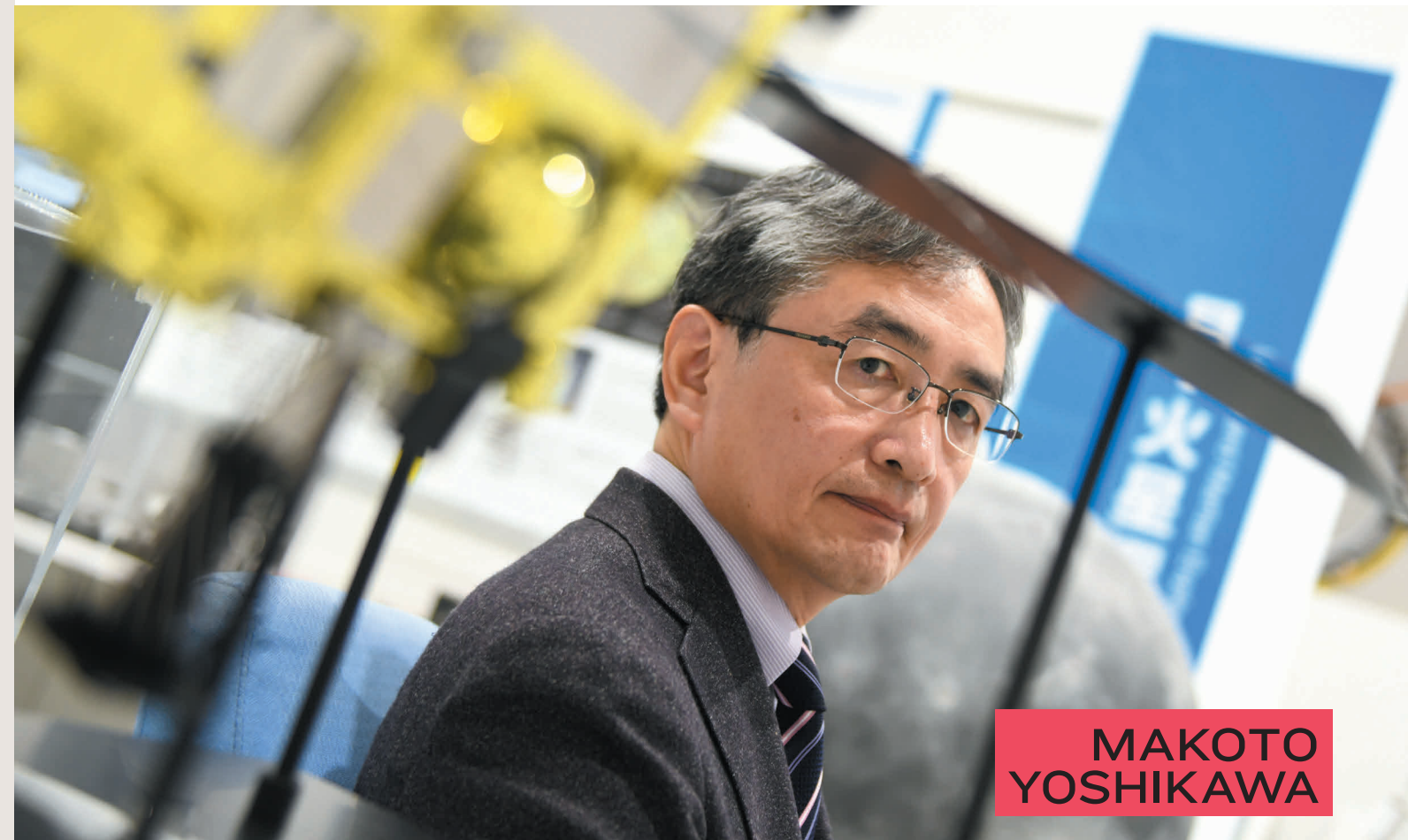
Those who have worked with Smits are not surprised by his ability to get consensus on controversial policies. "Robert-Jan has a fantastic memory, of people, events, documents, policies. His networking capacity is spectacular," says Helga Nowotny, a former president of the ERC.

Smits' short tenure as open-access tsar is almost over. Next year, he will leave to become chair of the Eindhoven University of Technology in his native Netherlands. "It's time for me to leave the commission at what I consider my height," he says. ■



ROBERT-JAN SMITS

ARTUR ERANOSIAN/EUROPEAN COMMISSION



MAKOTO YOSHIKAWA

ASTEROID HUNTER

An astronomer led a daring mission to collect samples from a rock in space.

BY DAVIDE CASTELVECCHI

In June 2018, astronomer Makoto Yoshikawa stayed up around the clock as the space mission he was leading zeroed in on its quarry — a dumpling-shaped rock called Ryugu. In a delicate manoeuvre after a journey of more than three years, the Hayabusa2 spacecraft fired its thrusters so that it moved in synchrony with the 1-kilometre-wide asteroid as they orbited the Sun together.

That task achieved, Yoshikawa and his team at the Japan Aerospace Exploration Agency (JAXA) moved on to the exploration phase. By early October, the craft had successfully dropped three small rovers onto Ryugu — providing the first close-ups of the asteroid.

Hayabusa2 faces a bigger test next year, when it will gently touch down on Ryugu and collect a sample. Any navigational imprecision could send it crashing against a boulder. In an even more daring manoeuvre, the craft will then shoot a projectile at the asteroid and analyse the material that gets kicked up. The probe is due to come back to Earth in 2020, carrying specimens that could shed light on the early stages of the Solar System's evolution.

Yoshikawa has been through nail-biters before. As a JAXA astronomer, he helped to mastermind two of the most spectacular rescue operations in the history of uncrewed space exploration.

The first mission to collect a sample from an asteroid, the original

Hayabusa, touched down on asteroid Itokawa in 2005. Soon afterwards, mission control lost touch with the craft. The team managed to restore communications and piloted Hayabusa back to Earth, despite having lost its main engine. The speeding craft burnt up during its re-entry, but its sample-return capsule was eventually recovered.

Then, in 2010, another JAXA probe, Akatsuki, had an engine malfunction as it tried to decelerate to enter into orbit around Venus. Akatsuki drifted away and went many times around the Sun until 2015, when it passed Venus again and the team managed to put it into orbit.

Some mishaps were inevitable, Yoshikawa says, given that Japan's space programme does not have a long tradition of deep-space exploration. "We need experience," he says. But Hayabusa2 has, so far, provided some redress for JAXA's historic ill-fortune.

Stephan Ulamec, a geophysicist at the German Aerospace Center in Cologne who had a leading role in developing one of the Hayabusa2 landers, MASCOT, says that risk-taking and the ability to learn from failures set Japanese space endeavours apart from more-cautious — and better-funded — agencies in the West. "They have a tendency to do bold missions, to take risks NASA would not," he says.

Yoshikawa has the rare ability to lead a collaboration of many different laboratories without having a big ego, and that has been key to the success of these missions, says Aurélie Moussi, an astrophysicist at the French space agency CNES in Toulouse and a co-project manager for MASCOT. "He is the kindest scientist I've ever worked with," she says.

Yoshikawa has had an interest in asteroids ever since he was a child and read *The Little Prince* — a 1943 novella that features a boy who lives on an asteroid and visits Earth. Asteroids are potential menaces that need to be kept track of — but they also hold the secrets to the Solar System, and are a possible source of materials to mine for future space exploration, Yoshikawa says.

"Asteroids are very small objects in the Universe — but very important for the future life of humans." ■

COMMENT

EXPLORATION Seafarers' journals are a rich record of discoveries **p.340**



CRISPR Time to redefine misleading meanings in genome editing **p.345**

OUTBREAKS WHO drafts code on pathogen sequence sharing **p.345**

OBITUARY Aaron Klug, electron-tomography Nobel laureate, remembered **p.346**

REPORTERS/EYEVINE



Women in a hospital ward with malaria bed nets in Bunia, the Democratic Republic of the Congo.

Vaccine candidates for poor nations are going to waste

Promising immunizations for diseases that affect mostly people in low- and middle-income countries need help getting to market, urge **David C. Kaslow** and colleagues.

Some 240 vaccine candidates are currently in the development pipeline for diseases such as malaria, tuberculosis and pneumonia — conditions that predominantly affect people in low- and middle-income countries (LMICs). Just two that made it through the pipeline in recent years are widely used in these nations: a conjugate vaccine for meningitis serogroup A diseases (see ‘Game changers’) and a vaccine against Japanese encephalitis virus.

Drug developers must clear many barriers to get a vaccine licensed so it can be given to millions of people — whether in routine immunizations or during a disease outbreak. The initial phases of development

entail evaluating candidates identified from basic research in proof-of-principle clinical trials, usually involving tens or hundreds of people (often called ‘the valley of death’ because so many candidates fail at this stage owing to a lack of resources; see *Nature* **453**, 840–842; 2008). Next, developers must invest in a manufacturing facility and test candidates in trials involving several thousand to tens of thousands of people (phase III). If regulators approve the vaccine for sale, vaccine manufacturers must then monitor safety in populations that have been vaccinated (known as post-marketing evaluation or phase IV studies).

Over the past decade, billions of

dollars have been ploughed into academic laboratories, biotechnology firms and pharmaceutical companies to help them through the first phase of vaccine development for diseases that mainly affect emerging economies. The money has come from many organizations, including the US National Institutes of Health (NIH), the European Union, the Wellcome Trust in London and the Bill & Melinda Gates Foundation in Seattle, Washington. Numerous candidates are now in proof-of-principle trials, largely thanks to these investments and to advances in technology that span genomics to immunology. Several candidates are ready for late-stage clinical trials. ►

GAME CHANGERS

The wider benefits of vaccination

Health workers in Uganda preparing for a meningitis vaccination campaign.

Vaccines are one of the most effective means of improving public health, technically and economically.

They benefit societies as well as individuals. They disrupt pathogen transmission, promote school attendance and attainment, and enable greater labour participation and productivity. In some instances, they have limited the spread of antimicrobial resistance, or even (in the case of smallpox) eradicated a disease⁷.

Between 2010 and 2017, the vaccination

of more than 300 million people in sub-Saharan Africa with meningococcal serogroup A conjugate vaccine has almost eliminated meningitis A^{8,9}. This disease type was causing hundreds of thousands to fall sick and, during the worst epidemic in 1996–97, resulted in more than 25,000 deaths. It has also removed a major cause of poverty in the region. Meningitis A diseases took up 50% of families' annual incomes, starting a spiral of poverty from which it was difficult to recover¹⁰. *D.C.K. et al.*

▶ Yet much of this promising pipeline could go to waste. No single organization or group is striving to support the formidably challenging second phase of vaccine development for diseases that mainly affect emerging economies.

LONG ROAD

Taking a vaccine candidate from a discovery at the lab bench to widespread deployment is complex, lengthy and expensive.

Take the RTS,S vaccine for malaria. In 1967, investigators showed that mice could be protected against rodent malaria by injecting them with a partially inactivated form of the malaria parasite *Plasmodium*¹. In 1983, researchers identified the gene encoding a protein on the surface of *Plasmodium*². This protein is recognized by antibodies that protect rhesus macaques from *Plasmodium* infection³. More than a decade later, studies revealed that a recombinant form of this protein could protect six out of every seven adult human volunteers from infection⁴. Then, in 2004, these findings were confirmed in a study involving more than 2,000 children from Mozambique⁵. From its

formal start in the mid-1980s, this first phase of development alone took about 20 years and cost around US\$200 million.

Beginning in 2009, phase III clinical studies were conducted in children from seven countries in sub-Saharan Africa⁶. The trials were run through a collaboration between the drug manufacturer GlaxoSmithKline (where R.R. is chief scientist) and the PATH Malaria Vaccine Initiative (MVI), headquartered in Seattle, Washington (D.K. was also involved in these studies). MVI received grant funding from the Bill & Melinda Gates Foundation.

In 2015, the vaccine was approved for market by the European Medicines Agency, on one condition: that the developers conduct post-marketing evaluations. But before a vaccine can be given to millions in LMICs, it must be recommended by the World Health Organization (WHO) and approved through the WHO prequalification process. Unexpectedly for the developers, in October 2015, the WHO's Strategic Advisory Group of Experts on Immunization and the Malaria Policy Advisory Committee called for the vaccine to be given on a pilot basis in just

three countries: Ghana, Kenya and Malawi.

The post-marketing studies are scheduled to begin next year. But the follow-up pilot studies are a massive undertaking: the relevant infrastructure must be set up in the selected countries. That includes training enough people to collect data in the field, as well as establishing systems to collect, store and analyse the data.

So far, the RTS,S vaccine programme has cost more than \$700 million. Several hundred million dollars more will be needed for a manufacturing plant (see 'Long and costly').

MARKET FAILURE

Who is going to cover the costs for other vaccine candidates in the pipeline?

We are most concerned about those that could have a significant impact on public health in LMICs, but that offer no or limited economic return for developers (see 'Spreading the cost').

Since 2000, the number of large companies developing and manufacturing vaccines has fallen from nine to four. (Manufacturers in emerging economies have tended to focus on producing vaccines that have already been developed in high-income countries.) This means that the burden of developing candidates for LMICs is now spread across fewer firms.

Some pharmaceutical giants, such as GlaxoSmithKline and Sanofi, have contributed for decades to the late-stage development of vaccines targeting diseases that mainly (or solely) affect LMICs. But even these firms are becoming more reluctant to invest as candidates are discovered at a faster rate, and as more of those are for diseases that affect only LMICs. As disease outbreaks become more frequent, such companies are similarly becoming averse to reallocating resources and disrupting daily operations to respond to public-health emergencies.

Some vaccine candidates in the current pipeline, such as those for Group A streptococcus, Group B streptococcus and tuberculosis, could have public and private markets in middle-income countries such as China. Yet developing them is a high-risk option. Various factors make it difficult for companies to predict the return on their investments. These can include: knowledge gaps about a candidate's mechanism of action, and an incomplete understanding of the impact of a disease — for instance, on the number of people infected, how many deaths it causes and the effect on a country's economy.

Sixty per cent of the vaccine candidates identified for diseases that affect LMICs target HIV, malaria, tuberculosis and pneumonia — diseases that are a much bigger problem in LMICs than in high-income countries. And nearly 90% of these candidates are in phase II clinical trials (involving hundreds or a few thousand people) or in

earlier stages of development.

These candidates could transform public health, but they are not a compelling business case for developers. As long as no organization plans and supports the development of this class of vaccine from beginning to end, it could be many decades before they improve global health and well-being. They could even languish altogether.

A WAY FORWARD

So what should be done? In our view, the main stakeholders must come together to define a new path for the sustainable development of vaccines that are socially justified but that have no business case, an uncertain one, or that require considerable public funding to reach the clinic.

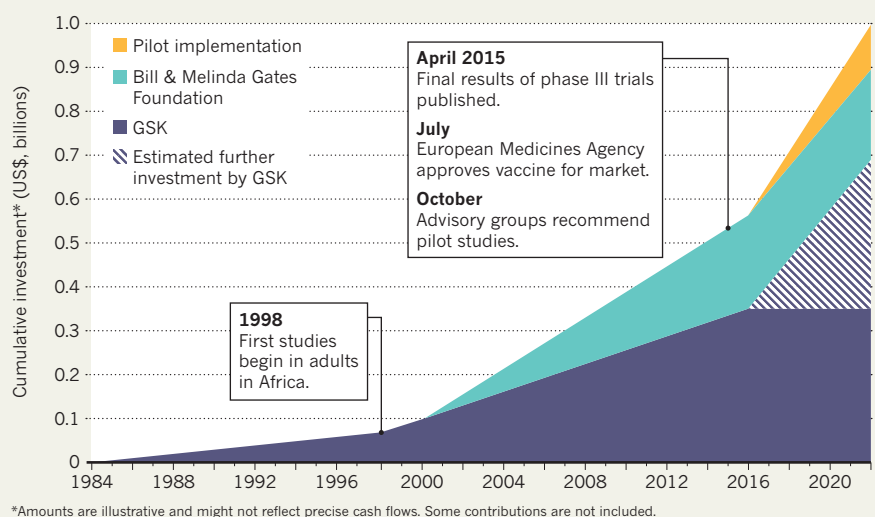
Participants should include funding agencies such as the Wellcome Trust, the Bill & Melinda Gates Foundation and the NIH; early-development partners such as PATH and the International Vaccine Institute in Seoul; vaccine manufacturers from wealthy and emerging economies; and organizations that recommend and procure vaccines for poor countries, such as the WHO and Gavi, the Vaccine Alliance.

The first aim should be to assess all the candidates in the pipeline. To ensure the best use of limited resources, the focus must be the most favourable candidates that face uncertain business cases. Those that are unlikely to have a significant impact socially and economically must be deprioritized.

Ultimately, the stakeholders must specify which organization, or alliance, should drive development for the prioritized vaccines from beginning to end, and what role each partner should have. They must also lay out the practical steps that are most likely to lead to the successful late-stage development and use of these vaccines — including schemes for resourcing.

LONG AND COSTLY

The drug giant GlaxoSmithKline (GSK) and others have spent nearly US\$1 billion on developing the RTS,S malaria vaccine. Millions more must be invested before it can be implemented at scale.



Money is the main limiting factor. In principle, subsidies from governments, such as those of the G20 countries, and philanthropic organizations such as the Bill & Melinda Gates Foundation, could remedy the market failure threatening vaccine development for LMICs. Gavi provides one form of subsidy (see 'Spreading the cost'). Support to develop vaccines or to make them available during epidemics is also provided by public organizations, such as the Coalition for Epidemic Preparedness Innovations in Oslo and the Biomedical Advanced Research and Development Authority, part of the US Department of Health and Human Services.

Such schemes need to be expanded and rethought to give vaccine developers more certainty and upfront financial backing. For instance, Gavi could commit to purchasing a vaccine before it has been developed, on the

condition that the developers meet certain regulatory milestones. At present, the alliance buys vaccines to distribute to LMICs after they have been licensed or recommended by the WHO for general use.

Regulation is another hurdle. Stakeholders should define clear pathways. They should negotiate more alignment between the various organizations involved in planning, development and oversight. Finally, they should identify what infrastructure and human capacity are needed to ensure that a reliable supply of vaccines can be provided long-term to the people who need them.

Only with this kind of leadership will the global community secure vaccines for some of the world's most debilitating diseases. ■

David C. Kaslow is vice-president for essential medicines at PATH, Seattle, Washington, USA. **Steve Black, David E. Bloom, Mahima Datla, David Salisbury, Rino Rappuoli.**
e-mail: dkaslow@path.org

SPREADING THE COST

How to fund vaccine development

The private sector might be able to advance some vaccine candidates for diseases that affect both emerging economies and wealthy nations.

Drug firms have provided many vaccines at a reduced cost to low- and middle-income countries (LMICs) — such as for poliovirus, hepatitis B and pneumococcus — after recovering their research and development investments through sales in wealthy nations. For instance, in the United States alone, the net economic benefit of the oral polio vaccine, licensed in 1961, was estimated to be more than US\$180 billion by 2006 (ref. 11).

Gaps in resourcing can limit vaccine

uptake in LMICs, even for diseases that affect high-income countries. The *Haemophilus influenzae* type b vaccine was adopted in Gambia, Kenya, Cuba and Nicaragua 12–16 years after it was incorporated into national US immunization programmes. This was largely because of the absence of a global agency able to purchase vaccines for the countries that couldn't afford them, and because suppliers were reluctant to invest in the manufacturing capacity needed for uncertain returns.

Various schemes have now been established, such as Gavi, the Vaccine Alliance, to help make existing vaccines available to people in LMICs. [D.C.K. et al.](#)

1. Nussenzweig, R. S., Vanderberg, J., Most, H. & Orton, C. *Nature* **216**, 160–162 (1967).
2. Ellis, J. et al. *Nature* **302**, 536–538 (1983).
3. Cochrane, A. H., Santoro, F., Nussenzweig, V., Gwatz, R. W. & Nussenzweig, R. S. *Proc. Natl Acad. Sci. USA* **79**, 5651–5655 (1982).
4. Stoute, J. A. et al. *N. Engl. J. Med.* **336**, 86–91 (1997).
5. Alonso, P. L. et al. *Lancet* **364**, 1411–1420 (2004).
6. RTS,S Clinical Trials Partnership. *Lancet* **386**, 31–45 (2015).
7. Bloom, D. E., Fan, Y. Y. & Sevilla, J. P. *Sci. Transl. Med.* **10**, eaa2345 (2018).
8. Trotter, C. L. et al. *Lancet Infect. Dis.* **17**, 867–872 (2017).
9. Mustapha, M. M. & Harrison, L. H. *Hum. Vaccin. Immunother.* **14**, 1107–1115 (2018).
10. Roberts, L. *Science* **320**, 1710–1715 (2008).
11. Thompson, K. M. et al. *Risk Anal.* **26**, 1571–1580 (2006).

S.B., D.E.B., M.D., D.S. & R.R. declare competing financial interests; see go.nature.com/2uhcjkw.



Captain James Cook's ships *Resolution* and *Discovery* beat through ice off Alaska in 1788. Watercolour by John Webber.

EXPLORATION

Centuries of science afloat

In explorers' journals, **Huw Lewis-Jones** found a rich record of discoveries made in cramped cabins, on open decks and on forays to shore to bring back proofs of marvels.

Samuel Johnson once remarked that “being in a ship is being in a jail, with the added chance of being drowned”. Hence the eighteenth-century lexicographer’s admiration for explorers: “The adventurer upon unknown coasts, and the describer of distant regions” is to be welcomed, he declared, because they “enlarge our knowledge”.

And enlarge knowledge they did, from

geography, oceanography and astronomy to meteorology, botany and zoology. Navigation was one of the greatest scientific challenges of Johnson’s time. Mapping had progressed steadily from the thirteenth century, when Italian merchant-venturers had developed the earliest portolan pilot charts of the Mediterranean, using compass directions and observations to locate harbours. The Atlantic voyages of European mariners to

the Americas, India and the Spice Islands — now the Maluku Islands of Indonesia — demanded re-engineered ships and growing expertise in celestial navigation. By the eighteenth century, some ships had become mobile labs in which instruments from sextants to chronometers were tested and improved, and ever more accurate sea charts plotted.

Beyond advances born of the need to stay on course, many seafarers kept journals,

LEBRECHT MUSIC & ARTS/ALAMY



recording minute observations of sea life, coastlines and curious natural phenomena. As I discovered while doing research for my forthcoming book *The Sea Journal*, these documents show science conducted *in situ* — in cramped cabins, on open deck and on exploratory forays from ship to shore. From the work of Venetian scholar Antonio Pigafetta in the sixteenth century to that of the first woman to circumnavigate the globe, botanist Jeanne Baret, in the eighteenth, they form an immensely valuable archive.

Nascent navies led the way in the newly important business of charting coasts and oceans for commerce and strategy. Navigators followed routes into the Pacific Ocean discovered by fifteenth-century Portuguese explorer Ferdinand Magellan, and later pushed south, searching for the coast of a suspected continent, or north into the Arctic maze of ice and islands. As familiarity with sea routes improved, explorers increasingly

used ships to get to the start of journeys inland. In the case of Antarctica (where I'm heading as I write this), they were sailing off the edge of the chart. The continent was not seen until 1820, and no one overwintered on its frozen landmass until 1899, when an expedition led by Norwegian explorer Carsten Borchgrevink from the ship *Southern Cross* survived through the darkness in a hut at Cape Adare. The more well-known expeditions that followed in the wake of whalers and sealers, such as those of Robert Falcon Scott, took observational science to the very edge of things.

SPONTANEITY AND RISK

'Explore' comes from the Latin *explorare*, to investigate. Many seafarers' journals also contain the word *adventure*, from the Old French *aventure*, 'to happen by chance', and the Latin *adventura*, a thing 'about to happen.' That mix of spontaneity, apprehension and risk lies at the heart of exploration: seafarers set forth to venture, to hazard, to bring back proofs of marvels. Pioneering natural observers on the move, they deployed all their skills and instrumentation to probe the unknown.

One of their great tools was draughting, long valued at sea for recording unfamiliar coastlines. The skill became important in sea officers' formal education. Some ships had official artists, and surviving sketchbooks speak of their achievements on course, or of being locked in floes, adrift, becalmed or waiting for whales to surface. (Sigismund Bacstrom, a little-known German surgeon with a passion for alchemy, spent almost five years on a voyage around the world, surviving stranding, a mutiny and time in a Mauritius jail before returning to London in 1795.)

Early maritime journals bear witness to the beginnings of scientific disciplines such as meteorology, oceanography and the discovery of species. And as established scientists increasingly took to ships, the art of recording found new levels of accuracy and precision.

British mariner James Cook, a veritable astronaut of the Enlightenment, set out on his voyages to the Pacific 250 years ago with well-chosen crews, many talented with brush and pen. Among them were father and son Johann and Georg Forster, as well as Sydney Parkinson and William Hodges. The *Endeavour* expedition changed our understanding of the cosmos, as astronomers used its observations of the 1769 transit of Venus across the Sun to measure Earth's distance from our star.

Cook was fond of the phrase "voyages of discovery" to describe his travels. Yet this brilliant cartographer, who journeyed emotionally and intellectually into waters unknown to him, encountered people from cultures that had navigated, voyaged and built understanding of their own marine environment for centuries. Tahitians, Hawaiians and Maori people would place this enigmatic visitor on their own maps in ways he could neither



A devilfish painted by Georg Forster in the 1770s.

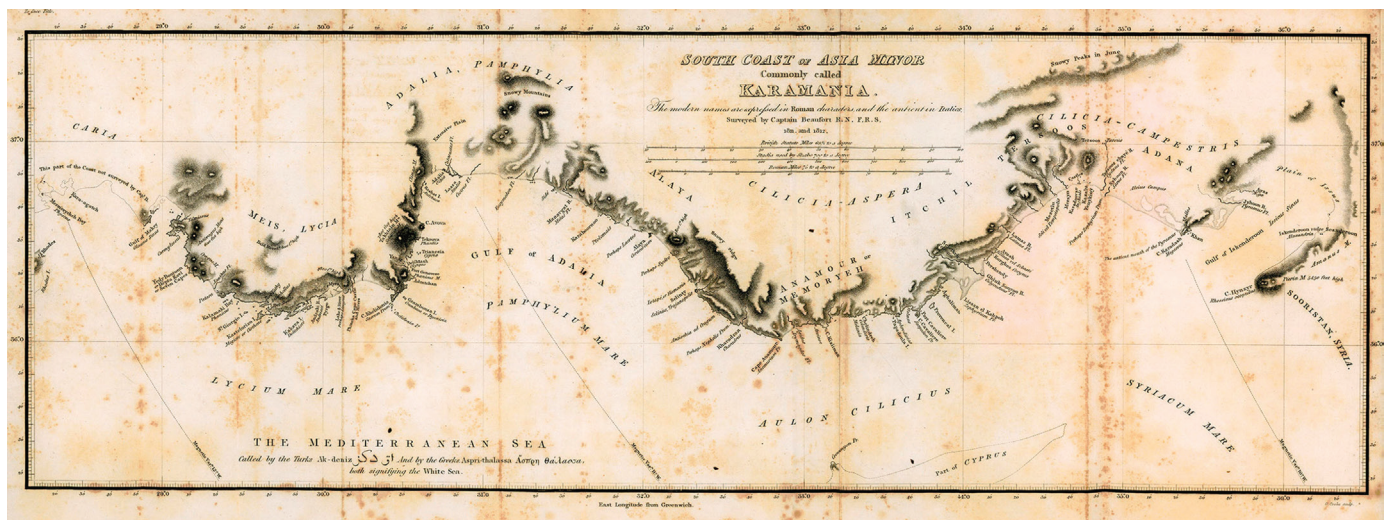
understand nor control. Consider Tahitian navigator Tupaia, whose knowledge of the Pacific's 'sea of islands' hugely assisted Cook. For generations of Polynesian voyagers, the sea was not so much an obstacle as a way.

Cook's exploits, charts, life and violent death in Hawaii have given him equal celebrity and notoriety. Less known is a young British vicar with a passion for mathematics, who held the fortunes of men, and nations, in his hands — by arbitrating a contest for new precision equipment. In 1763, Nevil Maskelyne was sent to Barbados to test a 'sea watch' devised by clockmaker John Harrison in a navigational arms race whose outcome would transform the world. Maskelyne's days were spent minutely inking out calculations as his ship lurched across the Atlantic.

His reward was to be made astronomer royal in 1765. He published the first volume of *The Nautical Almanac* two years later. This celebrated work contained a table of lunar distances for computing longitude, and was assembled with a team of human 'computers' — assistants versed in arithmetic, geometry, trigonometry and observational acuity.

MAPPING INFORMATION

A generation later, Irish hydrographer Francis Beaufort orchestrated one of the greatest mapping exercises ever attempted. With British merchant and naval fleets commanding the seas, it was Beaufort's job at the Hydrographic Office of the Admiralty to track information on which ships relied, from where to lay anchor to intelligence about fortification and trade. Officers' naval journals were crammed with such data, and Beaufort's Admiralty charts became the gold standard. In them, centuries of the art and science of seafaring are distilled into two dimensions and half a square metre.



An 1817 map of the south coast of Anatolia, surveyed by Francis Beaufort.

Along with devising an eponymous wind-force scale in 1805, Beaufort was a prime mover behind the Admiralty's *Manual of Scientific Enquiry*, first appearing in 1849. This contained instructions for observation in a dizzying range of new disciplines — a scientific A–Z from astronomy to zoology, by way of botany, geology, hydrography, magnetism, mineralogy, statistics and tides. Charles Darwin and botanist William Hooker contributed essays. From a time when British surveying vessels could be found in New Zealand, the Torres Strait and the Arctic, this monumental work is a snapshot of UK domination of the seas. Its instructions for seafarers were clear: “Let him then acquire the habit of never quitting his ship without his note-book and pencil, and his pocket-compass.”

As Victorian science gave way to the twentieth century's game-changing discoveries — from continental drift to other galaxies — British rule of the oceans ebbed. The journal habit did not. The 1915 *Nature Notes for Ocean Voyagers*, co-authored by sea captain Alfred Carpenter, was filled with his “personal observations upon life in the ‘vast deep’” — one of the earliest spotters' guides to marine wildlife. He was assisted by scholar-sailor David Wilson Barker, who had ferried passengers to gold-rush Australia and served on iron ships laying submarine cables on the ocean floor. Barker was a keen observer of seabird migrations and the shapes of waves, and found time to sketch the aurora australis and discover remarkable plankton blooms off the coast of Ecuador. An author of seamanship and navigation manuals, he also penned the 1918 *Things A Sailor Needs to Know*.

By the 1930s, US oceanographer and zoologist William Beebe was enrapturing the world with radio commentary delivered from a research submarine,

the ‘bathysphere’, hundreds of metres down in Bermuda waters. Beebe headed the Department of Tropical Research at the New York Zoological Society and encouraged female scientists to work with him. Of the many who joined his expeditions, Gloria Hollister is particularly interesting. A trained zoologist and cancer researcher, she applied to work with Beebe in 1928, as he was looking for a professional naturalist, with skills in dissection, for an expedition to Bermuda.

Hollister became invaluable as an experienced ichthyologist, and she made research descents in the bathysphere. Her observational skills, and talent with paint and ink, turned Beebe's scientific discoveries into true works of art.

Hollister eventually led three expeditions herself. In Trinidad, she explored the Arma Gorge and studied the oilbird (*Steatornis caripensis*), the world's only nocturnal flying, fruit-eating bird. In 1936, she embarked on an expedition to Guyana's Kaieteur Falls, trekking through more than 300 kilometres of dense tropical jungle. With her team, she discovered species and brought back specimens, including the curiously reptilian-looking stinkbird, or hoatzin (*Opisthocomus hoazin*). In the early 1950s, she co-created the Mianus River Gorge Conservation Committee, saving this habitat in Bedford, New York, from development. Throughout, Hollister, like Beebe, kept extensive logs and notes. Her career began, as many other field scientists' did, with discoveries made from the shifting deck of a ship.

In the pages of rare journals and sea logs, private diaries and cloth-bound sketchbooks, there is much for historians of science to discover. Observation remains at the heart of scientific effort, and worlds are still there for the exploring — from the microbiome to the sea bed and into deep space. As the Renaissance humanist Petrarch put it, people go forth to behold “the mighty surge of the sea ... the inexhaustible ocean, and the paths of the stars”. And in so doing, they “lose themselves in wonderment”. ■

Huw Lewis-Jones is an environmental historian, expedition guide and senior lecturer at Falmouth University, UK. He wrote *Explorers' Sketchbooks* with Kari Herbert, and his monograph *The Sea Journal* will be published in 2019. Twitter: @polarworld



Botanist Jeanne Baret, disguised as a man to join an expedition.

Correspondence

CRISPR twins: China academy responds

As representatives of the Committee of Genome Editing of the Genetics Society of China and of the Chinese Society for Stem Cell Research, we were shocked by He Jiankui's claims last month that twin girls were born from embryos that were gene-edited for HIV resistance (*Nature* **563**, 607–608; 2018). Such work would violate the current code of conduct from China's ministry of health, as well as internationally accepted ethical guidelines (see go.nature.com/2erqwpcc).

The consensus of the international scientific community, including Chinese researchers in genome editing, is that engineering the human germline for reproductive purposes should be forbidden until the scientific issues have been resolved and there is broad social agreement. China has clear regulations specifying that human embryos with genetic modifications cannot be implanted, in agreement with regulations adopted worldwide.

Genome editing in somatic cells holds promise for treating many genetic diseases. This powerful technology must not be abused or allowed to undermine the trust of regulators and the public in responsible scientific research.

Wensheng Wei* *Peking University, Beijing, China.*

**On behalf of 5 correspondents (see go.nature.com/2gflbtff for full list). wswei@pku.edu.cn*

CRISPR twins: what does 'editing' mean?

In view of the far-reaching implications of the birth of allegedly 'gene-edited' twin girls announced by Chinese researcher He Jiankui last month (*Nature* **563**, 607–608; 2018), we urgently need to revisit the use of the term.

It is ten years since the concept of gene editing took off (see, for

example, E. E. Perez *et al.* *Nature Biotechnol.* **26**, 808–816; 2008). This was used to describe just about any DNA modification by exogenous nuclease systems. It now makes more sense to apply it only to deliberate, precise alterations to DNA sequences. Sequences modified haphazardly by cells after the introduction of CRISPR would then be classified simply as random mutations and not as 'gene edits'.

This is not just a matter of semantics (see also M. O'Keefe *et al.* *Am. J. Bioeth.* **15**, 3–10; 2015). Characterizing He's claimed mutations to the *CCR5* gene as 'edits' misleads the public by implying that they were planned and applied with accuracy. It seems, however, that they were the result of random insertions and deletions of DNA. Exaggerating the precision of the process is harmful — in part, because it downplays the potential biological risks associated with random gene mutations in the germline.

Overall, a more-precise definition of genome editing will be helpful in the human reproductive context — in the event of more 'CRISPR babies' — and for broader CRISPR-related applications.

Paul Knoepfler *University of California, Davis, USA.*
knoepfler@ucdavis.edu

WHO code on free outbreak data

To understand and control disease outbreaks, researchers need free access to the genetic sequences of pathogenic organisms as soon as they are ready (N. L. Yozwiak *et al.* *Nature* **518**, 477–479; 2015).

The World Health Organization (WHO) is proposing a code of conduct for the public release of pathogen genomic sequences at the time of disease outbreaks. By making it easier to share the benefits rapidly and equitably, the code will help public-health authorities, product developers and researchers to

collaborate more effectively, and from a position of mutual trust.

The WHO code of conduct is based on consultations with stakeholders and on lessons from recent outbreaks. Crucially, all parties must recognize the importance of early pathogen sequencing and early public sharing of data and benefits, before and during outbreaks. Sequence sharing before publication should become standard; secondary users and data providers need to collaborate on reports of sequence analyses; and international partners should support local sequencing efforts and develop a sequence-analysis network. Exploring different models for sharing sequence data will allow for the preferences of data providers.

The draft code of conduct is available at go.nature.com/2bbllkts and the deadline for commenting is 28 January 2019.
Vasee Moorthy, Peter Salama, Soumya Swaminathan *WHO, Geneva, Switzerland.*
moorthyv@who.int

Rallying cry to halt biodiversity loss

Writing on behalf of the authors of the biodiversity section of the latest Global Environment Outlook (GEO-6) from the United Nations Environment Programme, to be released in March 2019 (see go.nature.com/2b9fp9o), we are concerned about your discussion on the progress of the IPBES assessment (see *Nature* **560**, 423–425; 2018). It risks diverting attention away from the scientific consensus on the perilous status and trends of biodiversity worldwide (see, for example, go.nature.com/2rttvwn).

GEO-6 indicates that policy responses have so far been insufficient to reduce or reverse biodiversity decline. Debate on best practice for conserving biodiversity is crucial. In our view, a 'conservation triage' approach must not prioritize reactive responses to environmental

pressures at the expense of reducing those pressures. Empirical evidence indicates that land sparing benefits biodiversity more than land sharing does, yet the 'half-Earth' concept — setting aside half the Earth for biodiversity — remains controversial. Indigenous people and local communities should not be overlooked. They can offer bottom-up and innovative solutions for protecting biodiversity.

We do not yet know whether we have entered a sixth mass extinction or whether there are planetary boundaries that could define a safe Earth system for people. Meanwhile, GEO-6 reinforces the stark message that the health of the planet and its people depend absolutely on biodiversity.

Jonathan Davies *University of British Columbia, Canada.*

Peter Stoett *University of Ontario Institute of Technology, Canada.*
j.davies@ubc.ca

King Faisal prize a Nobel harbinger?

The prestigious King Faisal International Prize for medicine (kingfaisalprize.org) was awarded in January to James Allison, who later shared this year's Nobel Prize in Medicine or Physiology (*Nature* **562**, 20–21; 2018). Gregory Winter, one of the three 2018 Nobel laureates in chemistry (*Nature* **562**, 176; 2018), won the King Faisal prize in 1995.

Like the Nobels, the King Faisal prize still has a long way to go towards rewarding women's contributions equitably. Out of the 113 winners of the King Faisal prizes in science and medicine since 1982, 4 were women (3.54%). This is marginally higher than for Nobels: just 20 women (3.29%) are named among the 607 Nobel prizewinners in the fields of science and medicine since 1901.
Sameen Ahmed Khan *Dhofar University, Salalah, Oman.*
rohelaakhan@yahoo.com

Aaron Klug

(1926–2018)

Crystallographer who won a Nobel prize for 3D imaging of viruses.

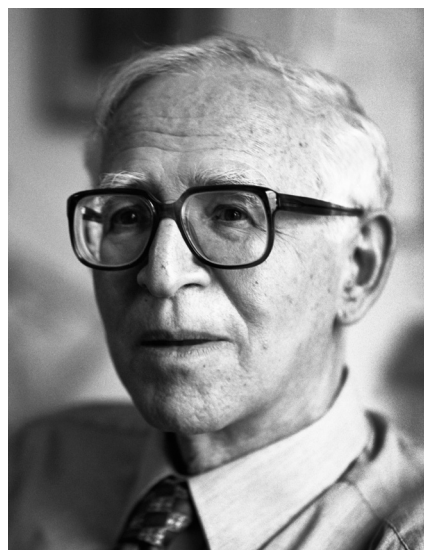
Aaron Klug solved the structures of macromolecules such as viruses, working with chemist Rosalind Franklin. He found the rules of the geometrical form of poliovirus and other spherical viruses. His invention of electron tomography, in which a 3D image of a virus is obtained from many electron micrographs, won him the Nobel Prize in Chemistry in 1982. He discovered proteins called zinc fingers that recognize DNA sequences and initiate the transcription of RNA — work that became the basis of gene therapy.

Between 1986 and 1996, he was director of the Medical Research Council's Laboratory of Molecular Biology (LMB) in Cambridge, UK. Here he was instrumental in the British part of the Human Genome Project, setting up the Sanger Centre in Hinxton with John Sulston as its director. He was president of the Royal Society from 1995 to 2000, charting a course through crises over bovine spongiform encephalopathy and transgenic crops.

In both jobs he galvanized action and reform. His mild manner belied a tough negotiator. He persuaded the Medical Research Council to allow grant recipients to hold patents. A decade later, income from licences was enough to make the LMB self-supporting. He was knighted in 1988 among myriad other honours. In 2013, the Ben-Gurion University of the Negev established the Aaron Klug Integrated Centre for Biomolecular Structure and Function.

Klug, who died on 20 November, was born in 1926 in Želva, Lithuania. His father was a ranch hand; his mother's family, the Silins, ran the shop in the shtetl. In 1928 Klug and his parents emigrated to Durban in South Africa. A precocious child, Klug was reading the newspaper aged three-and-a-half. At 15, he won a scholarship to study medicine at the University of the Witwatersrand in Johannesburg (in the same class as another future Nobel-prizewinner and LMB head, Sydney Brenner). Klug switched after a year to natural science and graduated with first-class honours in physics, chemistry and mathematics.

In 1946, Klug did a master's degree in the physics department at the University of Cape Town with Reginald James, who had worked with Lawrence Bragg, the founding father of X-ray crystallography. James had an enduring influence: Klug came to rely on his friendship and advice. In 1947, Klug published his first paper (A. Klug *Nature* **160**, 570; 1947). Two further papers in *Acta Crystallographica* on the structure of triphenylene ($C_{18}H_{12}$)



earned him a scholarship to Trinity College, Cambridge. In 1948, he married Liebe Bobrow, a dancer and musician. The next year they left for the United Kingdom.

In Cambridge in 1952, Klug completed a doctorate on the kinetics of steel formation in the lab of physicist Douglas Hartree. The year after, he did theoretical studies on the kinetics of oxygen uptake by the blood pigment haemoglobin with physiologist F. J. W. Roughton. This reawakened his interest in biological phenomena.

Then, in one of the great 'what ifs' of science history, Klug was denied a US visa — on the grounds that he had belonged to a student organization that the South African government deemed communist; McCarthyism was at its peak. So in 1953, he moved to London, to the lab of crystallographer J. D. Bernal at Birkbeck College.

Klug met Rosalind Franklin on a staircase there. It was an epiphany: his theoretical gifts combined perfectly with her technical skills. The two collaborated on the structure of tobacco mosaic virus. Their work earned them a grant from the US National Institutes of Health, a rare thing then for an organization outside the United States. After Franklin's premature death in 1958, Klug took over leadership of the group, working with structural biologist Donald Caspar on spherical viruses.

Such viruses have icosahedral symmetry. The maximum number of molecules that can be evenly assembled on the surface of an icosahedron is 60, but the number found in viruses is always much larger. Klug and Caspar introduced the idea of quasi-equivalence

in virus structures, which allows multiples of 60. This was done by inscribing triangles on the surfaces of an icosahedron, an idea borrowed from Buckminster Fuller. All known spherical viruses can be classified in this way.

In 1962, the group moved to the newly founded LMB. Klug and his long-term collaborator John Finch turned to electron microscopy to determine the triangulation number of several spherical viruses. The images were difficult to interpret because of the superposition of the front and back surfaces, which led some to doubt the Caspar–Klug rules. With biophysicist David DeRosier, Klug showed in 1968 how to calculate a 3D image by combining many 2D snapshots of the particle from different angles, deconstructing information in the images using mathematical manipulations called Fourier transforms. The 3D structures showed that Klug's interpretations were correct. (Computed tomography in medicine was later developed independently by Godfrey Hounsfield and Allan Cormack, a friend and colleague from Klug's Cape Town days.)

In 1985, Klug discovered modular proteins called zinc fingers. He realized it would be possible to string together many specific zinc fingers to recognize any DNA sequence, which inspired the design of synthetic zinc fingers to target a wide range of diseases. Later, he began working on neurodegeneration, noting the importance of the tau protein in Alzheimer's disease.

Aaron was a gifted teacher. He had a phenomenal memory and an encyclopaedic knowledge of most things, driven by unbridled curiosity. An admired theoretician, he was assisted by supportive experimentalists throughout his career. Friendly and caring, he inspired loyalty. As director of the LMB he was a valued boss, willing to lend a sympathetic ear. At the Royal Society, three vice-presidents wrote a letter expressing their admiration: "Aaron Klug brings to the Presidency intellectual rigour and integrity, penetrating insights and knowledge of a staggering array of fields, both scientific and cultural." ■

Kenneth Holmes was director of the Department of Biophysics at the Max Planck Institute for Medical Research, Heidelberg, Germany, from 1968 to 2003. He wrote *Aaron Klug: A Long Way from Durban* (Cambridge Univ. Press, 2017). He was introduced to Klug by Rosalind Franklin in 1955 at Birkbeck, and worked with him at the LMB. They were close friends and colleagues. e-mail: ken.holmes@mpimf-heidelberg.mpg.de

Tiny crystals, big potential

A method called microcrystal electron diffraction can rapidly image the structures of small molecules, including those found in mixtures. Will it usurp X-ray crystallography for determining small-molecule structures?

ALAN BROWN & JON CLARDY

For decades, X-ray crystallography has been the gold-standard method for determining the structures of small molecules. But two papers, one by Gruene *et al.*¹ in *Angewandte Chemie* and the other by Jones *et al.*² in *ACS Central Science*, illustrate the potential of using an electron cryomicroscopy (cryo-EM) method called microcrystal electron diffraction (MicroED) for this purpose. Using different laboratory set-ups, the two groups have produced strikingly concordant sets of results: high-resolution structures of a suite of small molecules, some of which were even solved from impure samples.

In X-ray crystallography, molecules are crystallized before being bombarded with X-rays. This is because the scattering of X-rays from single molecules is too weak to be used for structure determinations, whereas the regularly repeating molecules in a crystal lattice focus the scattering into stronger patterns that can be analysed to generate a 3D atomic model of the molecule. However, X-ray crystallography requires samples to be relatively pure. Moreover, growing crystals large enough to diffract X-rays that will produce a measurable signal is an artisanal skill, and a bottleneck for structure determination.

In MicroED, crystals are exposed to a focused electron beam in an electron microscope, rather than to X-rays. The technique was originally used to solve protein structures from microcrystals³, but the new papers demonstrate its potential for determining the structures of small organic molecules. Electrons interact with matter much more strongly than X-rays do, which makes it possible to use crystals just one-millionth of the size of the crystals used in X-ray diffraction. The crystals are rotated in the electron beam by tilting the microscope stage to create a series of diffraction patterns. Because the stage has a limited tilt range, the method often requires data from multiple crystals to be merged.

The stronger interaction of the irradiating electrons with the molecular target comes at a price: more damage is done to the sample than in X-ray crystallography. To reduce damage, MicroED uses a highly attenuated electron beam, and the crystals are frozen and imaged under cryogenic conditions. Freezing prolongs

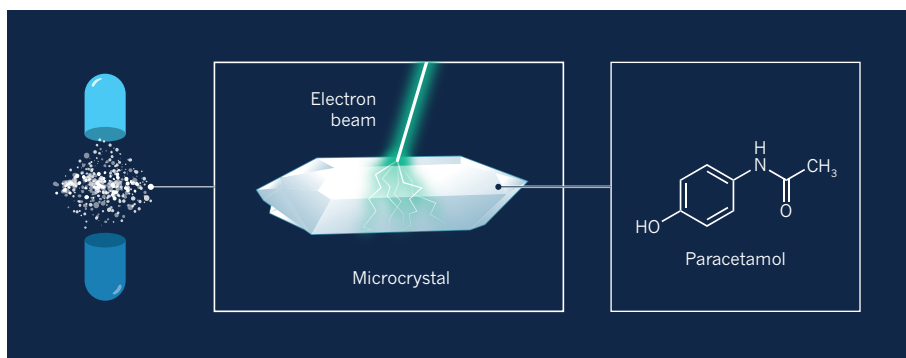


Figure 1 | Microcrystal electron diffraction (MicroED) applied to small molecules. Gruene *et al.*¹ and Jones *et al.*² have demonstrated that MicroED can determine the structures of small molecules in impure samples — a feat that could not be accomplished using X-ray crystallography. For example, Gruene and colleagues solved the structure of paracetamol from tablets of a cold medication that contained several biologically active and inactive ingredients. They focused an electron beam precisely at microcrystals of paracetamol in the medication, and then recorded and analysed the diffraction patterns to visualize the structure.

the life of the crystals, and thereby extends the range of samples that can be studied, compared with related electron-diffraction techniques that work at room temperature^{4,5} and are suitable only for highly stable crystals.

Jones and colleagues report that they were able to use MicroED to solve the structure of paracetamol from over-the-counter painkillers ground up using a mortar and pestle. Gruene and co-workers independently solved the same structure from tablets of a cold medication that contained several biologically active and inactive ingredients (Fig. 1). Both groups also solved other structures to illustrate the broad applicability of MicroED. One particularly impressive example reported by Jones *et al.* was the structure of thiostrepton, a complex peptide antibiotic that has a globular structure ten times larger than paracetamol's. In all cases, both groups used crystals that were orders of magnitude smaller than those required for X-ray crystallography.

The possibilities of MicroED are tantalizing. Using this technique, it should be possible to obtain the structures of compounds that cannot be coaxed to form large crystals. Additionally, imperfect or intergrown crystals could also be studied by either breaking them into smaller, regular pieces or directing the electron beam at small parts of the crystal.

However, the most exciting application would be to obtain structures of compounds

in tiny samples of mixtures. This should be possible because electron beams can be finely controlled and directed only at crystals, ignoring non-crystalline contaminants. For example, one could imagine using MicroED to study compounds extracted from natural sources, which are often obtained in extremely small quantities. Genome-based approaches are uncovering naturally occurring molecules at an ever-increasing rate⁶, the structural characterization of which has lagged behind their discovery.

It remains to be seen whether MicroED will have the same revolutionary impact on chemistry that other cryo-EM methods have had on other fields (particularly structural biology). The excitement that has been generated around MicroED should be tempered by the fact that several other methods — such as atomic-force microscopy⁷ and crystalline sponges⁸ — have previously been feted as replacements for X-ray crystallography, but have failed to live up to expectations for one reason or another.

One factor that might allow MicroED to succeed is that the electron microscopes and detectors needed for it are already found in most modern electron-microscopy facilities. The software used for data acquisition will also be familiar to many electron microscopists, and electron-diffraction data can be processed using the same software as that used

NEUROSCIENCE

to process X-ray-diffraction data. The entry barrier for any structural biologist or chemist wanting to use MicroED is therefore low. But chemists might need to compete for time on electron microscopes with their colleagues in biology departments.

Ultimately, the adoption of MicroED might depend on what percentage of small molecules are amenable to the technique. Previous work³ in which MicroED was used to solve protein structures from microcrystals suggests that there is no limitation on molecular size — it should work for everything from small organic molecules to large, multiprotein complexes. Nevertheless, MicroED does not remove the need for crystals, and not every molecule will crystallize. The technique is also unable to distinguish between mirror-image isomers of molecules, which is a drawback because such isomers can have very different biological properties.

Should the technique take off, the next step will be to develop electron microscopes specifically for small-molecule analysis. These microscopes would be the same as those currently used in structural biology, but would have detectors that are optimized for electron diffraction, and stages that have a greater tilt range and that can be more finely controlled. The development of systems for automated data collection and structure determination would allow the rapid, routine determination of structures from complex mixtures. Given that data collection is fast (a whole data set can be collected from one microcrystal in just three minutes), many thousands of crystals could be imaged from a single sample of material.

Small-molecule MicroED might also teach us a lot about how electrons interact with matter. Unlike X-rays, which interact only with electron clouds in molecules, electrons interact with both protons and electrons. Finally, knowledge gained from small-molecule structures solved at atomic resolution should help to improve the quality of all structures solved by cryo-EM methods, from small-molecule drugs to multiprotein biological machines. ■

Alan Brown and Jon Clardy are in the Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA. e-mails: alan_brown@hms.harvard.edu; jon_clardy@hms.harvard.edu

1. Gruene, T. *et al.* *Angew. Chemie Int. Edn* **57**, 16313–16317 (2018).
2. Jones, C. G. *et al.* *ACS Central Sci.* **4**, 1587–1592 (2018).
3. Shi, D., Nannenga, B. L., Iadanza, M. G. & Gonen, T. *eLife* **2**, e01345 (2013).
4. Mugnaioli, E., Gorelik, T. & Kolb, U. *Ultramicroscopy* **109**, 758–765 (2009).
5. Su, J. *et al.* *Microporous Mesoporous Matter* **189**, 115–125 (2014).
6. Zerikly, M. & Challis, G. L. *ChemBioChem* **10**, 625–633 (2009).
7. Gross, L., Mohn, F., Moll, N., Liljeroth, P. & Meyer, G. *Science* **325**, 1110–1114 (2009).
8. Inokuma, Y. *et al.* *Nature* **495**, 461–466 (2013).

Brain circuits of compulsive addiction

A study in mice identifies a brain adaptation that underlies the compulsive behaviour associated with drug addiction, and which might explain why some drug users behave compulsively whereas others do not. SEE ARTICLE P.366

PATRICIA JANAK

Drugs of abuse have complex pharmacological effects that trigger many changes in brain function. One of these effects, the direct or indirect activation of neurons that release the neurotransmitter dopamine, is common to all drugs of abuse and has long been assumed to contribute to the development of addiction. On page 366, Pascoli *et al.*¹ report on the neurobiological mechanisms induced by the repeated activation of dopamine neurons that might explain why some drug users seek reward despite facing negative consequences — a type of compulsive behaviour that is a defining feature of human addiction².

The authors took an optogenetics approach to mimic the activation of the brain's dopamine systems by drugs of abuse: they used laser light delivered through an optical fibre to activate dopamine neurons in the ventral tegmental area (VTA) of the brains of genetically engineered mice. The mice could directly stimulate these neurons themselves by pressing a lever, and performed this action avidly during a test

period of 40 minutes a day for almost 2 weeks.

On subsequent days, the mice received a brief electric shock to their feet on one-third of the lever-pressing occasions, at random. Their behaviour under this condition revealed an intriguing variability: 40% of the mice (termed renouncers) greatly reduced the frequency of lever-pressing when given foot shocks (Fig. 1a), whereas the remaining 60% (perseverers) were willing to receive painful punishment for the opportunity to self-stimulate their dopamine neurons (Fig. 1b). As some of these authors have previously shown³, the persevering mice provide a model for persistent drug use despite negative consequences, and parallel the subset of human drug users whose drug use becomes compulsive.

The authors next tried to determine what was different between the brains of perseverers and renouncers. They measured the activity of neurons connecting different brain areas in real time to determine which networks were active when mice pressed the lever. Communication between the orbitofrontal cortex (OFC), an area involved in decision-making, and the dorsal striatum, which is engaged in voluntary

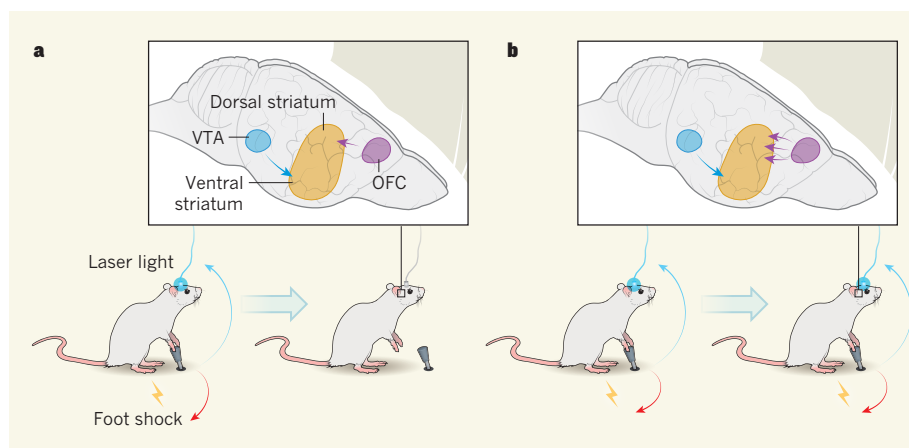


Figure 1 | Compulsive activation of dopamine neurons in the brain. In the study by Pascoli *et al.*¹, mice pressed a lever to activate dopamine-releasing neurons through the delivery of laser light conducted by an optical fibre. These neurons, which project from the ventral tegmental area (VTA) to the ventral striatum in the brain, are associated with reward. **a**, Some mice, termed renouncers, reduced the lever-pressing behaviour when it was associated with a painful electric shock to their feet. The strength of the connections between neurons of the orbitofrontal cortex (OFC) projecting to the dorsal striatum was low in these mice. **b**, Other mice, termed perseverers, continued to press the lever despite the punishment — a hallmark of compulsive behaviour. The neural connections between the OFC and the dorsal striatum were stronger in these mice than in renouncers. When the authors weakened these connections in persevering mice, the animals' compulsive behaviour decreased (not shown).

action, increased before lever-pressing in mice that were willing to obtain shocks along with dopamine self-stimulation. Optogenetic inhibition of this neural pathway turned persevering mice into renouncing mice. This finding shows that the increased activity of neurons projecting from the OFC to the dorsal striatum was necessary for this form of compulsive activation of dopamine neurons.

However, this behavioural switch was only temporary: when optogenetic inhibition was turned off, the compulsive behaviour resumed in persevering mice. The authors reasoned that long-lasting changes at the synapses — the junctions between neurons — that connect OFC and dorsal striatum neurons could arise as a result of the many days of self-stimulation of dopamine neurons. If these changes occurred only in persevering mice, this would explain their persistent compulsive behaviour.

If this hypothesis is true, the strength of synaptic connections between OFC and dorsal striatum neurons should be greater in perseverers than in renouncers, enabling better activation of dorsal striatum neurons by OFC neurons. Indeed, Pascoli *et al.* went on to show that the strength of the synapses between OFC neurons and dorsal striatum neurons had increased in persevering mice (Fig. 1). Renouncers, along with mice that had never been exposed to the experimental setup and mice that received shocks but were not allowed to use the lever, all showed low synaptic strength between OFC and dorsal striatum neurons.

Remarkably, the authors found that compulsive behaviour could be suppressed or induced by respectively decreasing or increasing the strength of this neural connection. Weakening of the synaptic connections between the OFC and the dorsal striatum in persevering mice reduced their willingness to self-stimulate in the face of a possible foot shock. Conversely, renouncers could be turned into perseverers by increasing the strength of these synaptic connections. In contrast to the temporary reversal observed after optogenetic inhibition of OFC neurons projecting to the dorsal striatum, these changes in synaptic strength induced a behavioural switch that persisted for six days.

Pascoli *et al.* have discovered a neuroadaptation that allows mice to override a painful stimulus to continue activating their dopamine neurons. The chronic consumption of drugs of abuse in humans leads to repeated activation of the same dopamine-reinforcement circuit, so a similar neuroadaptation might cause them to continue taking drugs despite the negative consequences. To test this proposition, we should determine whether changes in the strength of the connections between OFC and dorsal striatum neurons mediate compulsive behaviour in mice pressing a lever to receive cocaine, amphetamines or opioids in the face of a possible foot shock.

Does the optogenetic stimulation of dopamine neurons accurately mimic the activation

of dopamine neurons by drugs of abuse? There are obvious differences between quickly switching a laser on and off during optogenetic stimulation and the slower onset and longer duration of drug action. Nevertheless, the authors previously showed⁴ that cocaine intake and optogenetic activation induce almost identical adaptations in dopamine neurons and their immediate downstream targets, providing a strong rationale for the experimental approach used in the current study.

Why does the self-stimulation of dopamine neurons lead to compulsive behaviour in only a subset of individuals? Persevering and renouncing mice self-stimulated for approximately the same time and with a similar number of events before foot-shock punishments began, yet the brains of the two groups seem to have changed in divergent ways. The VTA dopamine neurons stimulated by the mice do not connect directly to the OFC or the dorsal striatum, so the link between these regions must involve multiple synaptic connections. A multisynaptic route through which the activation of VTA dopamine neurons might cause changes in the dorsal striatum has previously been described⁵, and has been proposed to underlie transitions from non-compulsive to compulsive drug-taking^{6,7}. Pre-existing differences in this multisynaptic circuit might explain why compulsive behaviour, and the related changes in synaptic

connections, occur in only some mice.

Synaptic changes can last for days, years or even a lifetime. Might the changes discovered by Pascoli *et al.* form the basis of an enduring behavioural change that is a hallmark of drug addiction? Resolving this question will require experimental evidence that drug self-administration despite negative consequences occurs through strengthening of the connections between the OFC and the dorsal striatum, and that it is indeed the activation of dopamine systems that sets in motion a chain of neural events that culminates in compulsive drug-taking. ■

Patricia Janak is in the Departments of Neuroscience and Psychological & Brain Sciences, Johns Hopkins University, Baltimore, Maryland 21218, USA.
e-mail: patricia.janak@jhu.edu

1. Pascoli, V. *et al.* *Nature* **564**, 366–371 (2018).
2. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders* 5th edn (Am. Psychiatr. Assoc., 2013).
3. Pascoli, P., Terrier, J., Hiver, A. & Lüscher, C. *Neuron* **88**, 1054–1066 (2015).
4. Brown, M. T. C., Korn, C. & Lüscher, C. *Channels* **5**, 461–463 (2011).
5. Haber, S. N., Fudge, J. L. & McFarland, N. R. *J. Neurosci.* **20**, 2369–2382 (2000).
6. Everitt, B. J. & Robbins, T. W. *Annu. Rev. Psychol.* **67**, 23–50 (2016).
7. Keiflin, R. & Janak, P. H. *Neuron* **88**, 247–263 (2015).

CONDENSED-MATTER PHYSICS

Elusive torque sensed by liquid crystals

Almost half a century ago, it was predicted that the confinement of quantum fluctuations could induce mechanical rotation — the Casimir torque. This prediction has now been confirmed using liquid crystals. SEE LETTER P.386

SLOBODAN ŽUMER

Quantum physics tells us that empty space is filled with fluctuating electromagnetic fields. If two metal plates are positioned close to each other, the quantum fluctuations between the plates differ from those outside the plates, producing a force that pushes the plates closer together. This phenomenon is known as the Casimir effect. In 1972, it was suggested¹ that quantum fluctuations could also generate a turning effect, called a torque, if the metal plates were replaced by materials that are optically anisotropic — that is, their optical properties, sensed by a light beam, depend on the beam's direction. On page 386, Somers *et al.*² report experimental evidence for this Casimir torque through the twisting of liquid crystals. The

discovery paves the way for the development of complex micrometre- and nanometre-scale mechanical devices.

Following the prediction of the Casimir effect between two ideal metal plates³, the concept was extended to real materials, such as conventional metals and electrical insulators known as dielectrics⁴. The Casimir effect can be explained by a restriction in the quantum and thermal fluctuations that can exist between the boundaries of two materials, leading to a weak attractive force. This force is maximal if the confining boundaries are identical, and is smaller — or can even be repulsive — if the boundaries differ in their electrical properties or shape⁵.

Because of the weakness of the Casimir force and its strong dependence on confining boundaries, it was nearly 50 years before solid experimental confirmation of the Casimir

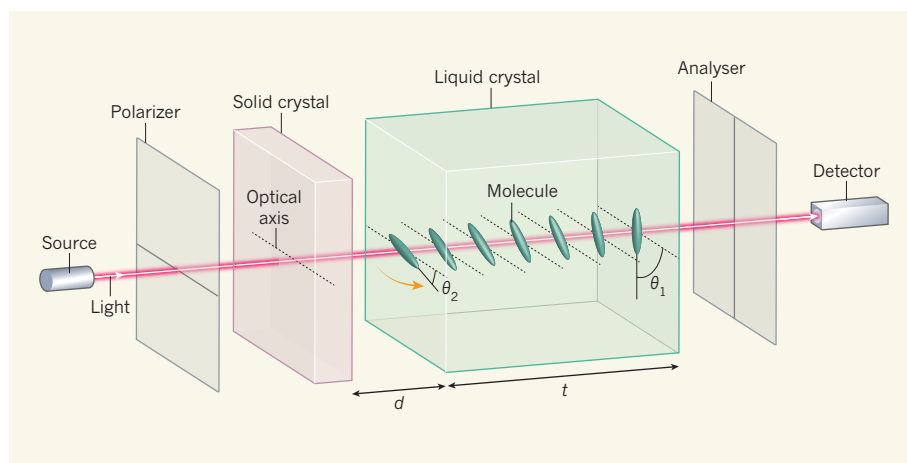


Figure 1 | Measurement of the Casimir torque. Somers *et al.*² have demonstrated that a solid crystal can cause the average orientation of molecules in a ‘nematic’ liquid crystal to twist (from angle θ_1 to angle θ_2), owing to a phenomenon known as the Casimir torque (orange arrow). The authors detected light that passed from a source through a polarizer, the two crystals and a second polarizer called an analyser. The thickness (t) of the liquid crystal was much larger than both the distance (d) between the crystals and the wavelength of the light. The authors selected a particular value for θ_1 and the orientations of the polarizers, and varied the solid crystal’s optical axis — a direction in the crystal along which the speed of propagating light is independent of the light’s polarization. They determined θ_2 from the intensity of the transmitted light, and measured the Casimir torque using the value of θ_2 and the known elastic properties of the liquid crystal.

effect was realized⁶. Concurrently, it became clear that the concept of fluctuation-induced interactions, arising from confined fluctuating electromagnetic fields, could be generalized to other confined fields in different areas of physics. Examples of such fields include sound waves and capillary waves (ripples).

Other fluidic systems that have fluctuation-induced interactions include ‘critical’ fluids⁷ and liquid crystals⁸. In the latter, the molecular ordering associated with different crystal phases — such as cholesteric, nematic and smectic phases — and, in particular, the anchoring of liquid crystals to surfaces can lead to diverse behaviour of Casimir-like forces⁹. It is also worth mentioning examples beyond condensed-matter systems, such as confined gravitational waves¹⁰.

The realization that altering the properties of the confining boundaries changes the ensuing Casimir force was naturally followed by the idea that breaking the symmetry of physical properties of confining surfaces or objects could lead to the emergence of the Casimir torque⁵. More than four decades ago, a theory was developed¹ to describe the Casimir torque between two solid, optically anisotropic crystals that are in close proximity to each other. This theory was subsequently improved and extended to more-complex systems such as layered media¹¹, and many experiments were proposed. Nevertheless, it is only now, with the work of Somers and colleagues, that the existence of the Casimir torque has been convincingly proved.

The authors’ groundbreaking experiment was based on the idea¹² of replacing one of the two solid crystals with a nematic liquid crystal. The liquid crystal was fixed on one side, and was exposed to the nearby solid crystal, which was free to rotate, on the other side (Fig. 1).

The liquid crystal had two roles: first as an optically anisotropic material, and second as a torque sensor.

The Casimir torque, although weak, forced the average orientation of molecules in the nematic liquid crystal along a direction that characterizes the solid crystal, known as its optical axis. This produced a twisted deformation that spread through the whole of the liquid crystal. Somers *et al.* detected the deformation by its effect on the intensity of light that was passed through a polarizer, the two crystals and a second polarizer called the analyser (Fig. 1). This

“The authors’ observation is a key contribution to fundamental physics that also has broad implications.”

Liquid crystals offer great potential because of their large response to weak external electric fields and other perturbations. Their best-known application is in liquid-crystal displays (LCDs). A less-known application is their use as sensor elements. Temperature sensors based on the temperature sensitivity of light reflection from cholesteric liquid crystals have been used for decades. By contrast, chemical sensors based on the adsorption of molecules on liquid-crystal surfaces, particularly for sensing biological molecules, have been developed only in the past decade. Somers and colleagues’ sensor for the Casimir torque is the latest example of a sensor in which a minuscule torque is determined using the elasticity of a liquid crystal, rather than by conventional mechanical means.

One remaining question is whether the measured Casimir torque could be enhanced. Much progress has been made in the optimization of nematic liquid crystals for displays and photonic applications, and so it would be worth exploring whether a liquid crystal is available that has more-optimal properties than has the crystal used by the authors. In addition to increasing the torque, such a liquid crystal would need to be more easily twisted than the one used here.

It is also possible that the torque sensor could be made from a cholesteric liquid crystal. Such a crystal selectively reflects circularly polarized light that has a polarization rotating in the same direction as the crystal’s intrinsically twisted structure. The Casimir torque would further twist, or untwist, the cholesteric crystal’s structure, affecting the selective light reflection in a way that should be detectable.

The successful observation of the Casimir torque by Somers *et al.* is a key contribution to fundamental physics that also has broad implications. The evolution of microscale mechanical and electromechanical devices has already reached to below the micrometre scale, and needs to take into account quantum phenomena and the effects of thermal and quantum fluctuations. Therefore, the development of nanoscale mechanical and electromechanical systems must take into account the Casimir force and torque, or even directly use these phenomena¹³.

Building on the current work, Casimir-like effects that can occur because of thermal fluctuations should be examined in many confined systems, including gases, conventional liquids, critical liquids, colloidal dispersions, polymers and liquid crystals. This could enable these systems to be used in complex micro- and nanoscale fluidic devices. ■

Slobodan Žumer is in the Department of Physics, Faculty of Mathematics and Physics, University of Ljubljana, and the Jožef Stefan Institute, 1000 Ljubljana, Slovenia.
e-mail: slobodan.zumer@fmf.uni-lj.si

- Parsegian, V. A. & Weiss, G. H. *J. Adhes.* **3**, 259–267 (1972).
- Somers, D. A. T., Garrett, J. L., Palm, K. J. & Munday, J. N. *Nature* **564**, 386–389 (2018).
- Casimir, H. B. G. *Proc. K. Ned. Akad. Wet.* **51**, 793–795 (1948).
- Lifshitz, E. M. *J. Exp. Theor. Phys. USSR* **29**, 94–110 (in Russian) (1955); *Sov. Phys.* **2**, 73–83 (1956).
- Woods, L. M. *et al. Rev. Mod. Phys.* **88**, 045003 (2016).
- Lamoreaux, S. K. *Phys. Rev. Lett.* **78**, 5–8 (1997).
- Hertlein, C., Helden, L., Gambassi, A., Dietrich, S. & Bechinger, C. *Nature* **451**, 172–175 (2008).
- Ajdari, A., Peliti, L. & Prost, J. *Phys. Rev. Lett.* **66**, 1481–1484 (1991).
- Ziherl, P., Podgornik, R. & Žumer, S. *Chem. Phys. Lett.* **95**, 99–104 (1998).
- Quach, J. Q. *Phys. Rev. Lett.* **114**, 081104 (2015).
- Lu, B.-S. & Podgornik, R. *J. Chem. Phys.* **145**, 044707 (2016).
- Smith, E. R. & Ninham, B. W. *Physica* **66**, 111–130 (1973).
- Capasso, F., Munday, J. N., Iannuzzi, D. & Chan, H. B. *IEEE J. Sel. Top. Quantum Electron.* **13**, 400–414 (2007).

MEDICAL RESEARCH

Success for cross-species heart transplants

A modified protocol has enabled baboons that received transplanted pig hearts to survive for more than six months. This improvement on previous efforts brings pig-to-human heart transplants a step closer. [SEE LETTER P.430](#)

CHRISTOPH KNOSALLA

Heart failure, in which the heart cannot pump blood around the body efficiently, is a problem of epic proportions. The number of adults living with heart failure in the United States is expected¹ to reach more than 8 million by 2030, and many of these people will die while waiting for a donor organ^{2,3}. One possible solution to this shortage is to use hearts from pig donors instead of from humans. But, so far, monkeys given transplanted pig hearts have not survived long-term, and so this approach has been deemed too risky to test in humans. On page 430, Längin *et al.*⁴ report modifications to a cross-species transplantation (xenotransplantation) approach that, for the first time, has enabled baboons that received genetically modified pig hearts to survive for more than six months.

In recent years, researchers have successfully transplanted kidneys from pigs into rhesus monkeys, with the transplants functioning for 435 days⁵. In addition, pig hearts transplanted into baboons that still had functioning hearts have survived for 945 days⁶. But in the latter case, the transplanted heart was not essential to the life of the recipient. Life-supporting

pig-to-baboon transplants have so far lasted only 57 days⁷.

Längin *et al.* set out to extend the survival of baboons receiving life-supporting heart transplants. They based their procedure on a previously described immunosuppression protocol⁶ that prevents the baboon immune system from rejecting the pig hearts, and used pigs that had been genetically modified to reduce interspecies immune reactions. A common criticism of xenotransplantation is that the immunosuppression protocols required are too toxic for use in humans. However, the protocol used by the authors seems to have been well tolerated by the baboons, with no major immunosuppression-related infections developing. Therefore, it might also be safe for use in humans, when and if xenotransplantation has advanced far enough to allow initial clinical trials.

The authors used an optimized process for preserving the pig hearts during transplantation. Typically, hearts are kept immersed in an ice-cold storage solution. However, the organ's tissue can be damaged when blood is recirculated through it. The researchers found that organ survival after transplantation could be improved by intermittently pumping

(perfusing) a blood-based, oxygenated solution containing nutrients and hormones through the hearts at 8°C during the procedure (Fig. 1).

This change improved short-term survival in four recipient baboons, but the animals died within 40 days owing to rapid, detrimental growth of the transplanted hearts. Längin and colleagues therefore modified the procedure to decrease this hypertrophy, and tested the optimized protocol in five more baboons. First, they reduced the baboons' blood pressure to match that of pigs. Second, they gave the baboons temsirolimus — a drug that combats heart overgrowth by stifling cell proliferation⁸. Third, they modified the standard hormone-treatment regimen. The steroid cortisone is typically given to transplant recipients to aid immunosuppression, but can cause heart overgrowth in newborn babies that receive stem-cell transplants⁹. The authors therefore tapered cortisone treatment much more quickly than they had for their first group of baboons, minimizing levels of the drug by three weeks after surgery.

Of the five baboons, one developed complications and was euthanized after 51 days. Two lived healthily for three months — the original designated endpoint of the experiment. The remaining two were allowed to survive for just over six months, before being euthanized.

The mechanisms underlying the consistency of transplant survival in Längin and colleagues' pig-to-baboon model need to be investigated. Nonetheless, the study's survival rate is impressive. A second finding also deserves recognition. In the past, all primates that received non-life-supporting heart xenotransplants and survived for more than three months developed a complication called consumptive coagulopathy, in which blood clotting increases in the microvessels of the transplanted heart^{6,10,11}. This condition results from a combination of intrinsic interspecies molecular incompatibility and natural immune responses. However, Längin *et al.* showed that consumptive coagulopathy could be prevented in their baboons by combining a genetic modification used in previous protocols — one that causes pigs to produce the human protein thrombomodulin, which reduces levels of clotting — with administration of temsirolimus (which inhibits aggregation of platelets in the blood).

Norman Shumway, the great pioneer of heart transplantation, is said to have believed, somewhat pessimistically, that xenotransplantation is the future of transplantation — and always will be. But the progress made by Längin and colleagues moves clinical heart xenotransplantation nearer to becoming a reality. As such, it is time to reconsider what preclinical results should be required before pig-to-human clinical trials can be initiated. Recommendations outlined by the International Society for Heart and Lung Transplantation in 2000 suggest that clinical trials might be considered once 60% of primates given life-supporting pig-heart

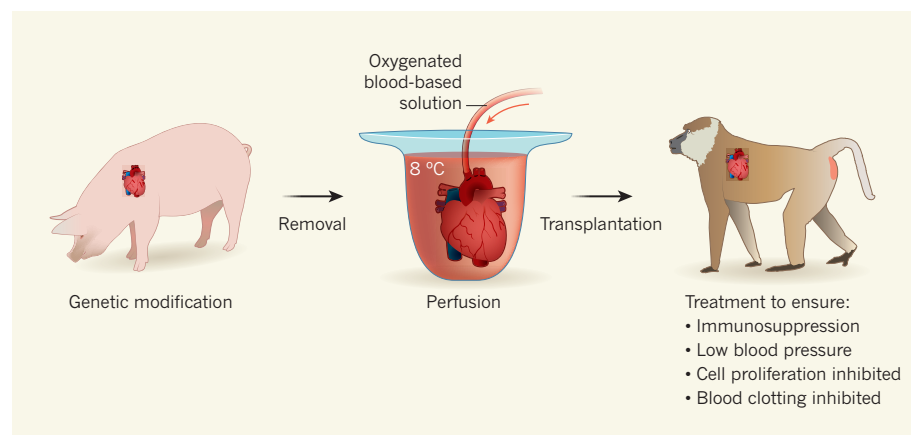


Figure 1 | Improving pig-to-primate heart transplants. Längin *et al.*⁴ took hearts from pigs that had been genetically modified to prevent the organs from triggering immune responses in baboons, and to prevent excessive blood clotting in the heart's blood vessels. To prevent the heart from becoming damaged through lack of blood during the transplant procedure, the authors intermittently pumped an oxygenated blood-based protective solution through the hearts (a technique called perfusion) at 8°C. The baboons that received the hearts had been subjected to a previously described treatment to ensure immunosuppression⁶. In addition, a newly modified combination of treatments lowered blood pressure, prevented blood clotting and blocked cell proliferation, preventing detrimental overgrowth of the heart following transplantation. Four of the five baboons survived for three months or longer.

transplants can survive for 3 months, with at least 10 animals surviving for this time frame, and with some indication that longer survival is possible¹². The current study goes some way to meeting these criteria. However, it seems likely that regulatory authorities such as the US Food and Drug Administration will require a longer period of follow-up and a greater percentage of successful experiments before permitting human trials.

In addition, other issues should be given attention before pig-to-human transplants become a reality. One such issue is the potential for pig viruses such as porcine endogenous retroviruses (PERVs) to be transmitted to humans. The risk of PERV-related complications is considered to be small¹³, but regulatory authorities worldwide still view the possibility with some caution. However, the genome-editing technology CRISPR–Cas has increased the speed with which pigs harbouring multiple genetic mutations can be generated, enabling researchers to produce live, healthy piglets in which PERVs have been deactivated¹⁴. This indicates one way of circumventing the risk of PERV transmission.

Another consideration is the fact that, in the past two decades, technology to improve blood circulation using mechanical support devices has evolved dramatically. These devices are used as a temporary fix while patients wait for a donor organ, but they can also be a permanent therapy for those with end-stage heart failure. The progress of this technology raises ethical questions regarding the use of pig hearts. For each patient, a case will have to be made for why a pig-heart transplant should be selected over mechanical support.

Regardless of the issues surrounding pig-to-human xenotransplantation, the blood-perfusion protocol exploited by Längin and colleagues could have a beneficial impact on human-to-human transplants. Cold static storage is still the standard for human organ transplants, but a blood-based solution could help to improve both short- and long-term results in the clinic. Moreover, it might allow the pool of donor hearts to be extended to include organs that are currently considered suboptimal because the donors are old or have an underlying condition that reduces the heart's ability to withstand the lack of a normal blood supply. ■

Christoph Knosalla is in the Department of Cardiothoracic and Vascular Surgery, German Heart Center Berlin, 13353 Berlin, Germany, and at the DZHK (German Centre for Cardiovascular Research), Partner Site Berlin, Berlin.
e-mail: knosalla@dhzb.de

4. Längin, M. *et al.* *Nature* **564**, 430–433 (2018).
5. Adams, A. B. *et al.* *Ann. Surg.* **268**, 564–573 (2018).
6. Mohiuddin, M. M. *et al.* *Nature Commun.* **7**, 11138 (2016).
7. Byrne, G. W., Du, Z., Sun, Z., Asmann, Y. W. & McGregor, C. G. A. *Xenotransplantation* **18**, 14–27 (2011).
8. Paoletti, E. *Transplantation* **102** (2S), S41–S43 (2018).
9. Lesnik, J. J., Singh, G. K., Balfour, I. C. & Wall, D. A.

- Bone Marrow Transplant* **27**, 1105–1108 (2001).
10. Kuwaki, K. *et al.* *Am. J. Transplant.* **4**, 363–372 (2004).
11. Kuwaki, K. *et al.* *Nature Med.* **11**, 29–31 (2005).
12. Cooper, D. K. C. *et al.* *J. Heart Lung Transplant.* **19**, 1125–1165 (2000).
13. Denner, J. *Science* **357**, 1238–1239 (2017).
14. Niu, D. *et al.* *Science* **357**, 1303–1307 (2017).

This article was published online on 5 December 2018.

ASTRONOMY

Abundant rare isotopes in a planetary nebula

Observations reveal that a particular planetary nebula — the ejected envelope of an old star — is unusually enriched in rare carbon, nitrogen and oxygen isotopes. The finding could help to explain the origins of these isotopes. [SEE LETTER P.378](#)

AMANDA KARAKAS

The origin of the chemical elements in the Universe is one of the most fascinating and enduring mysteries in astronomy. Progress so far has come from studies of stars, but here only elemental abundances can be determined reliably. Isotopic ratios are more difficult to obtain. On page 378, Schmidt *et al.*¹ study the composition of the young planetary nebula K4-47 — a glowing shell of gas and dust that formed from the outer layer of a Sun-like star and that was thrown off during the final stages of the star's evolution. The authors find that the nebula is unusually enriched in rare isotopes of carbon (¹³C), nitrogen (¹⁵N) and oxygen (¹⁷O). The measured composition of K4-47 shows that this object is more enriched in these isotopes than is almost any other nebula or star examined so far.

Why is Schmidt and colleagues' finding such a big deal? For one thing, it seems to suggest that stars similar to the Sun can make these rare isotopes — a result that was not expected. Computer simulations² of Sun-like stars have shown that they can be factories for carbon and nitrogen, but only in the form of the dominant isotopes ¹²C and ¹⁴N. Furthermore, theory² predicts that the rarer isotopes are not made inside stars that become planetary nebulae. What about direct observations of ageing Sun-like stars, as opposed to planetary nebulae? Such observations are difficult, but the available data^{3,4} mostly agree with theory, making K4-47 a particularly unusual object.

The only instance in which the isotopes ¹³C, ¹⁵N and ¹⁷O are synthesized at the same time is in explosions. CNO cycles are a collection of thermonuclear reactions that involve the capture of protons by isotopes of carbon, nitrogen

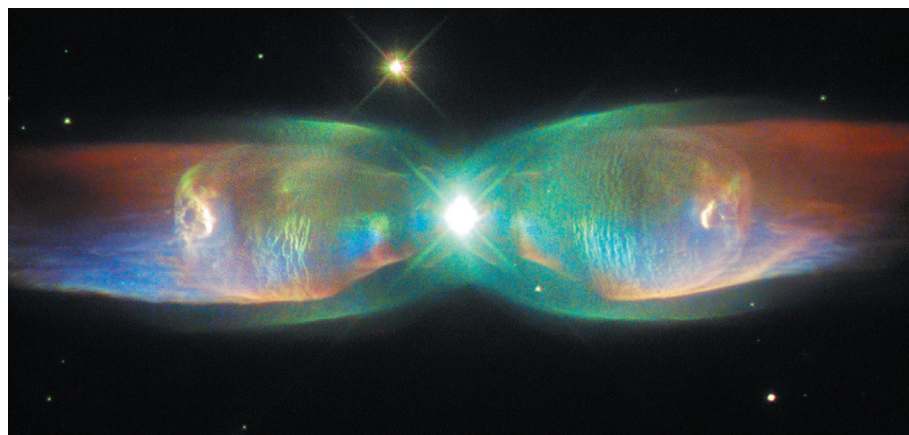


Figure 1 | A bipolar planetary nebula. A planetary nebula is a glowing shell of gas and dust that is ejected from a Sun-like star during the final stages of the star's evolution. Shown here is the planetary nebula M2-9 (also known as the Twin Jet nebula). Schmidt *et al.*¹ report observations of the planetary nebula K4-47 that, like M2-9, has an hourglass (bipolar) shape and highly collimated outflows of material. The authors find that K4-47 contains an unexpectedly high abundance of rare isotopes of carbon, nitrogen and oxygen.

1. Benjamin, E. J. *et al.* *Circulation* **137**, e67–e492 (2018).
2. Colvin, M. *et al.* *Am. J. Transplant.* **18** (Suppl. 1), 291–362 (2018).
3. Branger, P. & Samuel, U. (eds) *Annual Report 2017: Eurotransplant International Foundation*; available at <https://go.nature.com/2deatzh> (2017).

and oxygen. These reactions are the workhorse of stellar energy production but do not make much ^{13}C , ^{15}N and ^{17}O . Generating these isotopes requires conditions of high temperature and density, as well as plenty of protons. Such a mechanism is known as the hot CNO cycle. So far, the products of the hot CNO cycle have been found only in classical novae⁵ — nuclear explosions that occur in certain binary star systems.

So how can the presence of the rare isotopes in K4-47 be explained? One mechanism proposed by Schmidt *et al.* is that the progenitor of K4-47 underwent an explosive event called a helium-shell flash immediately before it became a planetary nebula. This is, in essence, a mixing event that causes hot material from the core of a star, rich in ^{12}C , to be moved to a cooler region, where hydrogen-fusion reactions are occurring. The mixing elevates the temperature of the cooler region, enabling reactions of the hot CNO cycle to proceed before the material is expelled to space.

Although the explosive nature of this scenario is unusual, similar mixing has previously been proposed to explain the composition of other chemically peculiar stars, such as Sakurai's object⁶ (also known as V4334 Sagittarii). Detailed computer simulations are needed to test this mechanism. If it can be verified, it will be evidence of previously unknown stellar behaviour that provides insight into how rare isotopes of common elements are generated.

But there are other possible explanations. The isotopic composition of K4-47 is similar to that of J-type carbon stars⁴, which have ratios of ^{12}C to ^{13}C of less than 15. The sequence of events that lead to a J-type star is unknown, and their existence is not predicted by the theory that describes the evolution of single stars. It has been suggested that J-type stars instead result from binary evolution⁷, in which two stars orbit each other and interact.

Such interactions have been proposed for all planetary nebulae that, like K4-47, have an hourglass (bipolar) shape and highly collimated outflows of material^{8,9} (Fig. 1). Observations show that the central stars of planetary nebulae are more likely to be binary stars than was previously thought, giving further credence to this idea. K4-47 could therefore be the product of an interaction or merger between two stars.

Alternatively, K4-47 might not be a planetary nebula at all. It has been speculated that it could be a planetary-nebula mimic, in which the extended nebula was ejected by a pair of interacting binary stars during an explosion¹⁰. The isotopic composition of K4-47 could be explained if the interaction of these stars resulted in an explosion akin to a classical nova that would allow for the hot CNO cycle. One prediction of this scenario is that gas would be ejected at high velocities. Has such ejection been observed?

Schmidt and colleagues say they have not seen these high-velocity outflows of material,

so they rule out a nova-like explosion as an explanation. But this finding is in contrast to previous studies that have observed high-velocity bullets of material ploughing through the surrounding medium^{11,12}. So who is right? Answering this question will require follow-up observations of K4-47 using astronomical instruments that can extract high-resolution spatial, dynamical and chemical information about the object.

Either way, K4-47, which is rich in the products normally associated with a nova but is embedded in something that looks like a planetary nebula, is one of the most isotopically unusual astronomical objects studied so far (along with CK Vulpeculae¹³). Detailed computer modelling and follow-up observations are required to tease out the true nature of the progenitor of K4-47. Such work could tell us something about how the rare isotopes of carbon, nitrogen and oxygen are made in stars. ■

Amanda Karakas is at the Monash Centre for Astrophysics, School of Physics and

Astronomy, Monash University, Victoria 3800, Australia.

e-mail: amanda.karakas@monash.edu

- Schmidt, D. R., Woolf, N. J., Zega, T. J. & Ziurys, L. M. *Nature* **564**, 378–381 (2018).
- Karakas, A. I. & Lattanzio, J. C. *Publ. Astron. Soc. Aust.* **31**, e030 (2014).
- Lambert, D. L., Gustafsson, B., Eriksson, K. & Hinkle, K. H. *Astrophys. J. Suppl. Ser.* **62**, 373–425 (1986).
- Abia, C. & Isern, J. *Mon. Not. R. Astron. Soc.* **289**, L11–L15 (1997).
- Gehrz, R. D., Truran, J. W., Williams, R. E. & Starrfield, S. *Publ. Astron. Soc. Pacif.* **110**, 3–26 (1998).
- Hervig, F. *et al. Astrophys. J.* **727**, 89 (2011).
- Zhang, X. & Jeffery, C. S. *Mon. Not. R. Astron. Soc.* **430**, 2113–2120 (2013).
- De Marco, O. *Publ. Astron. Soc. Pacif.* **121**, 316–342 (2009).
- Jones, D. & Boffin, H. M. J. *Nature Astron.* **1**, 0117 (2017).
- Corradi, R. L. M. *et al. Astrophys. J.* **535**, 823–832 (2000).
- Gonçalves, D. R. *et al. Mon. Not. R. Astron. Soc.* **355**, 37–43 (2004).
- Akras, S., Gonçalves, D. R. & Ramos-Larios, G. *Mon. Not. R. Astron. Soc.* **465**, 1289–1296 (2017).
- Kamiński, T. *et al. Astron. Astrophys.* **607**, A78 (2017).

NEURODEGENERATION

Amyloid- β ‘seeds’ in old growth-hormone vials

Some samples of human growth hormone used as therapy until the mid-1980s contain amyloid- β peptide and cause genetically modified mice to develop amyloid- β deposits in the brain. [SEE LETTER P.415](#)

TIEN-PHAT V. HUYNH & DAVID M. HOLTZMAN

In cerebral amyloid angiopathy (CAA) and Alzheimer's disease, insoluble aggregates of a peptide known as amyloid- β ($\text{A}\beta$) progressively build up in the spaces between cells, forming amyloid deposits. In Alzheimer's disease, these aggregates are found between neurons, whereas in CAA, a related but not always coexisting condition, they are found in the walls of brain blood vessels. $\text{A}\beta$ aggregates are thought to be early drivers of the pathological processes of CAA and Alzheimer's disease that culminate in neurodegeneration. In 2015, researchers reported evidence of early $\text{A}\beta$ pathology in the brains of some people with growth deficiency who had been treated with human growth hormone collected from pituitary glands at autopsy¹. This finding raised the possibility that $\text{A}\beta$ pathology might be transmissible between humans under certain conditions through contaminated brain-tissue derivatives. On page 415, Purro *et al.*² provide further support for this hypothesis.

From 1958 to 1985, approximately 30,000 children with growth deficiency were treated with cadaver-derived growth hormone

(c-hGH) worldwide³. In 1985, three recipients were found to have developed Creutzfeldt-Jakob disease (CJD), which is fatal. CJD belongs to a group of diseases known as transmissible spongiform encephalopathies, which are characterized by progressive and irreversible brain damage resulting from the accumulation of a misfolded form of a brain protein called prion protein. These abnormal prion proteins can themselves cause normal prion proteins to misfold, and thus spread the disease. Given the evidence that contaminated c-hGH had caused CJD, this type of treatment was quickly stopped and synthetic recombinant human growth hormone (rhGH) became the standard of care.

Alzheimer's disease is not a classic prion disease, but shares characteristics with this type of disorder. Misfolded, aggregated $\text{A}\beta$ peptides and tau proteins, which are toxic to neurons, are present in the brain as key components of Alzheimer's disease. Inoculation of minute amounts of misfolded $\text{A}\beta$ (known as $\text{A}\beta$ ‘seeds’) isolated from the brains of people with Alzheimer's disease can induce build-up of $\text{A}\beta$ deposits (called $\text{A}\beta$ plaques) in non-human primates⁴, and brain extracts from people or

mice that develop A β plaques can also cause accelerated plaque accumulation when given to genetically modified mice⁵.

The 2015 finding of A β plaques and CAA in the brains of seven of eight recipients of c-hGH therapy who had died of CJD further supported the idea that A β pathology can be transmitted through a prion-like mechanism¹ (Fig. 1). A β pathology is rarely found in young adults without genetic risk factors for Alzheimer's disease or CAA, so the findings suggested that the c-hGH used to treat the patients might have been contaminated with A β seeds in addition to misfolded prion proteins.

To provide more-direct evidence that the A β deposits found in these people resulted from A β -seed contamination, Purro and colleagues first tested whether A β was present in vials of c-hGH from batches that had been used to treat patients and that had been stored since the 1980s. Growth hormone is produced in the pituitary gland, a small structure found at the base of the brain. To obtain c-hGH, the pituitary glands from thousands of donors had been pooled and mixed, and the hormone had been chemically extracted using various preparation methods. Patients received c-hGH from multiple batches. However, all of those who were treated in the United Kingdom and developed CJD — 38 people by the year 2000 (ref. 6) — received injections from batches prepared using a method called the Hartree-modified Wilhelmi procedure (HWP).

Purro *et al.* detected A β in all c-hGH samples prepared using the HWP method, but not in those prepared using any of three other methods. Size-exclusion chromatography, a separation technique used in all non-HWP preparations, might have reduced contamination by A β peptides.

The authors went on to show that these HWP preparations of c-hGH possess A β -seeding ability by injecting them into mice genetically engineered to express human versions of A β (Fig. 1). Mice inoculated with HWP-prepared c-hGH developed markedly more A β plaques and CAA than did those inoculated with synthetic rhGH.

These results provide strong evidence that the A β pathology previously reported in people who died of CJD after receiving c-hGH¹ was indeed caused by their treatment. The data also corroborate previous studies in genetically modified mice demonstrating that misfolded A β can behave in a prion-like fashion⁵. Future studies should investigate the amount of A β these patients received over their treatment course, to try to determine the threshold of misfolded A β concentration required to transmit A β -plaque formation or CAA.

The c-hGH preparations shown in this study to induce A β pathology in mouse brains were injected directly into the brain, whereas the affected humans had received injections through other routes (intravenously or intramuscularly). Future studies in animals should assess whether the route of administration

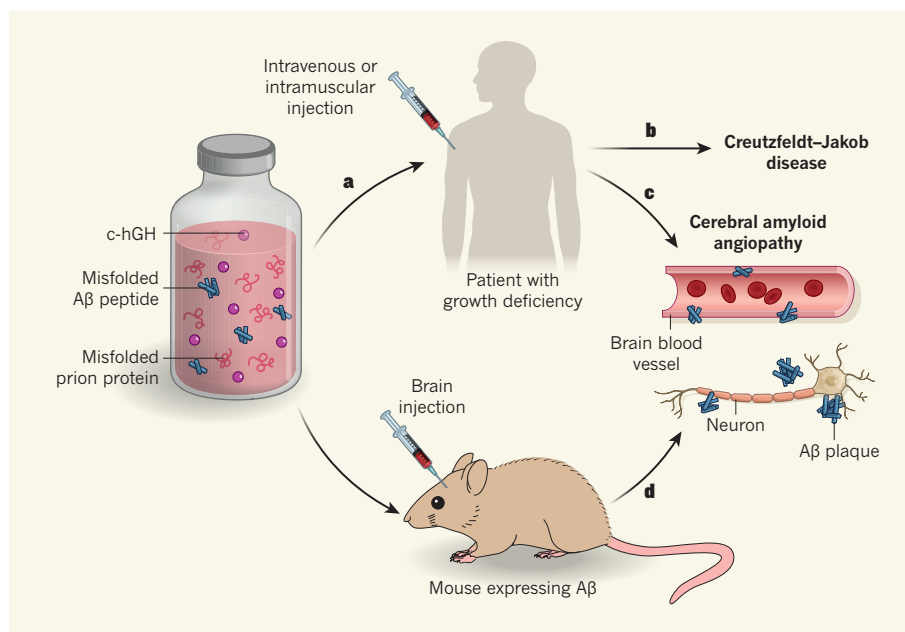


Figure 1 | Treatment-induced transmission of prion-like pathology. **a**, From 1958 to 1985, cadaver-derived human growth hormone (c-hGH) was used to treat growth deficiency. **b**, A few batches were contaminated with misfolded prion proteins from people with Creutzfeldt–Jakob disease (CJD) and caused c-hGH-treated patients to develop CJD. **c**, Intriguingly, some of these patients also had early signs of cerebral amyloid angiopathy (CAA), a condition that frequently co-occurs with Alzheimer's disease, and amyloid- β (A β) peptide aggregates (plaques) between neurons. This suggested that c-hGH vials were also contaminated with misfolded A β peptide (A β seeds), the main component of the pathological brain deposits that characterize Alzheimer's disease and CAA¹. Purro *et al.*² report that some of the c-hGH samples still in storage do indeed contain A β . **d**, The authors show that injection of c-hGH materials containing A β into the brain of A β -expressing mice leads to accumulation of A β , especially in brain blood vessels. These results demonstrate that A β seeds retain their pathological ability for a long time, and that they can potentially be transmitted through medical procedures.

influences the ability of material containing misfolded A β to cause brain A β pathology, and should investigate the minimum amount of material that has pathological effects.

The eight people with therapy-induced CJD who, with one exception, also had A β pathology had an incubation period from their last c-hGH treatment to CJD onset of 18.8–30.8 years¹. A β accumulation in people who develop dementia due to Alzheimer's disease is estimated to precede disease symptoms by 15–20 years^{7,8}. The seven people who developed A β pathology did not meet the full pathological criteria of Alzheimer's disease, and whether they would have developed the clinical manifestations of the disease had they not died of CJD is unclear.

Purro *et al.*² report that the HWP c-hGH batches also contained misfolded tau proteins. This study did not show evidence of tau-pathology transmission, and the earlier study of people with therapy-induced CJD did not detect misfolded tau proteins in their brains¹. Nevertheless, surveillance of surviving c-hGH recipients should continue to watch out for this possibility. Overall, treatment-related transmission of various brain pathologies cannot be ruled out.

Lastly, it is worth noting that the stored vials of c-hGH had been maintained at ambient temperature since the mid-1980s. Their ability to transmit A β pathology seen in this

study corroborates the idea that A β seeds are remarkably stable⁹. This property of A β seeds emphasizes the importance of not using biological material prepared from the human central nervous system for injection or transplantation into patients during neurosurgical or medical procedures, unless these materials are adequately screened or there is no other option. Similarly, it is crucial that surgical instruments that come into contact with the human brain are appropriately treated to remove seeds of misfolded forms of peptides and proteins such as A β , tau or prion protein. ■

Tien-Phat V. Huynh and David M. Holtzman are in the Department of Neurology, Washington University in St. Louis, St Louis, Missouri 63110, USA.
e-mail: holtzman@wustl.edu

1. Jaunmuktane, Z. *et al.* *Nature* **525**, 247–250 (2015).
2. Purro, S. A. *et al.* *Nature* **564**, 415–419 (2018).
3. Will, R. G. *Br. Med. Bull.* **66**, 255–265 (2003).
4. Baker, H. F., Ridley, R. M., Duchon, L. W., Crow, T. J. & Bruton, C. J. *Int. J. Exp. Pathol.* **74**, 441–454 (1993).
5. Meyer-Luehmann, M. *et al.* *Science* **313**, 1781–1784 (2006).
6. Swerdlow, A. J., Higgins, C. D., Adlard, P., Jones, M. E. & Preece, M. A. *Neurology* **61**, 783–791 (2003).
7. Fagan, A. M. *et al.* *Sci. Transl. Med.* **6**, 226ra230 (2014).
8. Jack, C. R. Jr & Holtzman, D. M. *Neuron* **80**, 1347–1358 (2013).
9. Ye, L. *et al.* *Nature Neurosci.* **18**, 1559–1561 (2015).

This article was published online on 13 December 2018.

365 DAYS:
the year in science

2018

EDITORS' CHOICE

Extracts from selected News & Views articles published this year.

NEUROSCIENCE

SENESCENCE MEDIATES NEURODEGENERATION

Jay Penney & Li-Huei Tsai (*Nature* 562, 503–504; 2018)

There is strong interest in understanding how neurodegeneration is affected by a cellular state called senescence, in which cells stop dividing, suppress intrinsic cell-death pathways and release pro-inflammatory molecules that can harm healthy neighbours. Bussian *et al.* focused on the neuronal protein tau, which, when in a mutated form dubbed tauP301S, causes the neurodegenerative condition frontotemporal dementia. The authors demonstrated that tauP301S expression in the neurons of mice can induce senescence in neuron-supporting cells called glia. In turn, senescent glia affect the ability of neurons to regulate tau phosphorylation and aggregation, ultimately promoting neuronal degeneration. The study adds to the growing body of evidence indicating that pharmacological removal of senescent cells could benefit people who have a wide range of conditions.

Original research: *Nature* 562, 578–582 (2018).

ORGANIC CHEMISTRY

A PRACTICAL ROUTE TO 3D MOLECULAR DIVERSITY

Wenbo Ye & Ang Li (*Nature* 560, 314–316; 2018)

Reactions known as cycloadditions construct molecules in a way that precisely controls the geometric arrangement of groups attached to carbon atoms. However, the scope of cycloadditions is limited to certain reactants, which has restricted their use for making libraries of compounds. Chen *et al.* report a strategy that combines cycloadditions with reactions known as carbon–carbon (C–C) cross couplings, to enable the modular and programmable preparation of cycloaddition-derived molecules. C–C cross-coupling reactions are often used to form bonds between carbon atoms that are already part of a carbon–carbon double bond. The use of such reactions to make compound libraries has resulted in a predominance of 2D molecules in compounds tested for drug discovery. 3D molecules tend to be rich in sp^3 carbons, which have the capacity to form four single bonds. Chen and colleagues' work will find numerous applications for synthesizing molecules rich in sp^3 carbons for drug discovery.

Original research: *Nature* 560, 350–354 (2018).



FIONA ROGERS/NPL

GENOMICS

NEWFOUND DIFFERENCES BETWEEN GREAT APES

Aylwyn Scally (*Nature* 559, 336–338; 2018)

The first great-ape genome projects used the human genome as a scaffold to help assemble genomic regions in which corresponding stretches of DNA lay in the same order and were present in a similar number of copies. But in regions where genome structure has evolved very differently in humans, the great-ape assemblies tended to be fragmented, leading to a deficit in our understanding of the genomic elements that make humans unique. Kronenberg *et al.* assembled high-quality genomes for a chimpanzee and an orangutan, along with two human genomes. They found more than 17,000 structural differences specific to humans, many of which disrupt genes. For instance, 41% of genes whose activity is suppressed in human progenitors for neurons and other cells in the brain's cortex are associated with a human-specific structural variant. This is consistent with structural genomic changes causing disruption or loss of gene function during great-ape evolution. Original research: *Science* 360, eaar6343 (2018).

OPTICAL PHYSICS

MIRRORS MADE OF A SINGLE ATOMIC LAYER

Kin Fai Mak & Jie Shan (*Nature* 556, 177–178; 2018)

The discovery of a single layer of carbon atoms, known as graphene, led to great interest in 2D materials. Whereas graphene is highly transparent to visible light, 2D materials that are highly reflective could be used as lightweight mirrors in optical or optoelectronic systems. The existence of such materials has been questioned, but, writing in *Physical Review Letters*, Back *et al.* and Scuri *et al.* report that single layers of molybdenum diselenide can have high levels of reflectance. Although the authors' near-perfect mirrors work only in light from a narrow range of the electromagnetic spectrum, the two studies open up intriguing possibilities for the fields of nanophotonics and quantum optics. The authors also demonstrate that the application of a voltage causes the mirrors to switch from being highly reflective to highly transparent.

Original research: *Phys. Rev. Lett.* 120, 037401 (2018); *Phys. Rev. Lett.* 120, 037402 (2018).

MICROBIOLOGY

BACTERIAL PERSISTENT CELLS TACKLED

Julian G. Hurdle & Aditi Deshpande (*Nature* 556, 40–41; 2018)

Chronic infections can be hard to treat because slow-growing bacteria known as persister cells are usually unharmed by antibiotics. The development of treatments for killing persister cells is needed to target the bacterium methicillin-resistant *Staphylococcus aureus* (MRSA), which is resistant to several antibiotics. Kim and colleagues searched for molecules that could protect worms from death mediated by MRSA infection. Of the compounds that conferred protection, the authors focused on two molecules known as retinoids. Prompt killing of cells occurred when the retinoids distorted the bacterial membrane's lipid bilayer. A major concern is how to optimize small molecules such as the retinoids to enable selectivity. The authors generated a compound that did not kill human cells, but did retain the ability to kill persister cells. Original research: *Nature* 556, 103–107 (2018).

SPACE PHYSICS

THE ORIGIN OF PULSATING AURORAS

Allison N. Jaynes (*Nature* 554, 302–303; 2018)

The Northern and Southern lights, also known as auroras, are as varied as the colours they display in the night sky. Discrete auroras (pictured) are the kind that typically grace our desktops and calendar covers, and are produced a few thousand kilometres above Earth's surface. By contrast, pulsating auroras are created tens of thousands of kilometres away, in the equatorial region of the magnetosphere — the area around Earth that is dominated by the planet's magnetic field. For decades, it has been suggested that pulsating auroras are the result of interactions between magnetospheric electrons and electromagnetic waves called chorus waves that send electrons careering towards Earth's atmosphere along magnetic-field lines. Kasahara *et al.* report direct evidence for this process using observations both from Earth's surface and from a spacecraft positioned on a field line.

Original research: *Nature* 554, 337–340 (2018).

ENVIRONMENTAL SCIENCE

THE FUTURE OF TIDAL WETLANDS IS IN OUR HANDS

Jonathan D. Woodruff (*Nature* 561, 183–185; 2018)

Coastal communities around the world depend on tidal marshes and mangroves for the diverse services these wetland systems provide. Such systems commonly reside just above mean sea level (pictured), putting them at risk of being drowned by rising sea levels. Schuerch *et al.* present global-scale modelling that suggests that tidal wetlands are less vulnerable to sea-level rise than was thought. Tidal wetlands can colonize new low-lying areas that become flooded by sea-level rise. The authors' results suggest that global tidal wetland loss will be 0–30% by 2100 for scenarios in which wetlands migrate only into sparsely populated regions, and if no action is taken to support or prevent migration. But wetland gains as high as 60% could be made when measures are taken that allow wetlands to migrate into more-populated areas.

Original research: *Nature* 561, 231–234 (2018).

GONZALO AZUMENDI/GETTY



DANITA DELIMONT/GETTY

READERS' CHOICE

We asked readers to vote for a News & Views article to be included as part of our round-up of the year. This is what they chose.

COMPUTATIONAL BIOCHEMISTRY

DESIGNER PROTEINS ACTIVATE FLUORESCENCE

Roberto A. Chica (*Nature* 561, 471–472; 2018)

If we could make proteins from scratch to bind any desired target molecule, it would open the door to a wide range of biotechnological applications. Dou *et al.* describe a computational method for designing proteins tailored to bind a small molecule of interest, and use it to make 'fluorescence-activating' proteins — biotechnological tools that have potential applications in biomedical research. The application of this method for designing 'β-barrel' proteins that bind small molecules is the first demonstration of the *de novo* design of both protein fold and function. Previous computational methods relied on building a binding cavity into a protein template found in nature. By contrast, Dou and co-workers have designed a β-barrel that has a shape distinct from those found in nature, and constructed a binding pocket that is specifically tailored to a target small molecule.

Original research: *Nature* 561, 485–491 (2018).

Soft-tissue evidence for homeothermy and crypsis in a Jurassic ichthyosaur

Johan Lindgren^{1*}, Peter Sjövall², Volker Thiel³, Wenxia Zheng⁴, Shosuke Ito⁵, Kazumasa Wakamatsu⁵, Rolf Hauff⁶, Benjamin P. Kear⁷, Anders Engdahl⁸, Carl Alwmark¹, Mats E. Eriksson¹, Martin Jarenmark¹, Sven Sachs⁹, Per E. Ahlberg^{10,11}, Federica Marone¹², Takeo Kuriyama^{13,14}, Ola Gustafsson¹⁵, Per Malmberg¹⁶, Aurélien Thomen¹⁷, Irene Rodríguez-Meizoso¹⁸, Per Uvdal¹⁹, Makoto Ojika²⁰ & Mary H. Schweitzer^{1,4,21}

Ichthyosaurs are extinct marine reptiles that display a notable external similarity to modern toothed whales. Here we show that this resemblance is more than skin deep. We apply a multidisciplinary experimental approach to characterize the cellular and molecular composition of integumental tissues in an exceptionally preserved specimen of the Early Jurassic ichthyosaur *Stenopterygius*. Our analyses recovered still-flexible remnants of the original scaleless skin, which comprises morphologically distinct epidermal and dermal layers. These are underlain by insulating blubber that would have augmented streamlining, buoyancy and homeothermy. Additionally, we identify endogenous proteinaceous and lipid constituents, together with keratinocytes and branched melanophores that contain eumelanin pigment. Distributional variation of melanophores across the body suggests countershading, possibly enhanced by physiological adjustments of colour to enable photoprotection, concealment and/or thermoregulation. Convergence of ichthyosaurs with extant marine amniotes thus extends to the ultrastructural and molecular levels, reflecting the omnipresent constraints of their shared adaptation to pelagic life.

With their dolphin-like external form, the Mesozoic ichthyosaurs are icons of evolution. Although finds from the 18th and 19th centuries led to their recognition as prehistoric oceangoing reptiles, it was not until the discovery of articulated skeletons with complete body outlines that the extent of this anatomical specialization towards an obligate marine existence was fully appreciated. These specimens, from the Holzmaden area of Germany, reveal a combination of paired flippers, a triangular dorsal fin and a lunate tail that would have enabled efficient hydrodynamic manoeuvring. Derived ichthyosaurs (parvipelvians) were thus notably similar in appearance to extant pelagic cruisers such as odontocete whales and lamnid sharks¹.

Despite this textbook example of evolutionary convergence, the nature of the residues that form the body contours—and the diagenetic processes that led to their preservation—remain incompletely understood. Various competing interpretations have suggested that these remnants comprise genuine soft parts (consisting either of carbonized and/or phosphatized integument, and superficial musculature or connective tissue^{2,3}), in situ transformed organic matter such as adipocere^{4,5}, or microorganismal replacement structures^{1,6}. Creases described as dermal tensile fibres have also been documented⁵, but alternatively dismissed as sedimentary cracks and tool marks⁷.

Here we use an integrated ultrastructural and molecular investigation of a chemically untreated specimen (MH 432; Urweltmuseum Hauff, Holzmaden, Germany) of the Early Jurassic (Toarcian) parvipelvic *Stenopterygius* to resolve these long-standing contentions. This fossil reveals endogenous cellular, sub-cellular and biomolecular constituents within relict skin and subcutaneous tissue (Figs. 1–5 and Extended Data

Figs. 1–7); we also document a possible internal organ trace, which we interpret as the liver (Extended Data Fig. 8). Finally, we provide comparative data obtained from the cutis and subcutis of modern marine tetrapods (Extended Data Fig. 9).

The fossil and experimental design

The soft-tissue residues of MH 432 comprise a buff-coloured to black coating that adheres to the outside of the postcranium, but also extends some distance beyond the periphery of the bones to produce a bedding-parallel outline of the animal (Fig. 1 and Extended Data Fig. 1a). A second, superficially amorphous material occurs as light beige patches external to this film in the abdominal region (Extended Data Fig. 1b), whereas fibrous matter (Extended Data Fig. 1c) and the red-brown liver trace are located inside the rib cage (Fig. 1 and Extended Data Fig. 1b).

We conducted examinations under regular (Fig. 1a and Extended Data Fig. 1a–c) and ultraviolet light (Extended Data Fig. 1d–m), as well as through synchrotron rapid-scanning X-ray fluorescence (Fig. 1c and Extended Data Fig. 1n, o). The resulting data enabled us to assess the elemental composition of the soft parts, and were also used to identify areas that were suitable for in-depth molecular and imaging investigations. Tissue and sediment samples (Fig. 1b) were collected with sterile instruments and distributed to multiple institutions for independent cross-referencing analyses. Some samples were subjected to chemical extraction, while others were first examined untreated, and then again after treatment with ethylenediaminetetraacetic acid (EDTA).

¹Department of Geology, Lund University, Lund, Sweden. ²RISE Research Institutes of Sweden, Chemistry and Materials, Borås, Sweden. ³Geobiology, Geoscience Centre, University of Göttingen, Göttingen, Germany. ⁴Department of Biological Sciences, North Carolina State University, Raleigh, NC, USA. ⁵Department of Chemistry, Fujita Health University School of Health Sciences, Toyoake, Japan. ⁶Urweltmuseum Hauff, Holzmaden, Germany. ⁷Museum of Evolution, Uppsala University, Uppsala, Sweden. ⁸MAX-IV laboratory, Lund University, Lund, Sweden. ⁹Naturkunde-Museum Bielefeld, Abteilung Geowissenschaften, Bielefeld, Germany. ¹⁰Science for Life Laboratory, Uppsala University, Uppsala, Sweden. ¹¹Department of Organismal Biology, Uppsala University, Uppsala, Sweden. ¹²Swiss Light Source, Paul Scherrer Institute, Villigen, Switzerland. ¹³Institute of Natural and Environmental Sciences, University of Hyogo, Hyogo, Japan. ¹⁴Wildlife Management Research Center, Hyogo, Japan. ¹⁵Department of Biology, Lund University, Lund, Sweden. ¹⁶Department of Chemistry and Chemical Engineering, Chalmers University of Technology, Gothenburg, Sweden. ¹⁷Department of Chemistry and Molecular Biology, University of Gothenburg, Gothenburg, Sweden. ¹⁸Centre for Analysis and Synthesis, Department of Chemistry, Lund University, Lund, Sweden. ¹⁹Chemical Physics, Department of Chemistry, Lund University, Lund, Sweden. ²⁰Department of Applied Biosciences, Graduate School of Bioagricultural Sciences, Nagoya University, Nagoya, Japan. ²¹North Carolina Museum of Natural Sciences, Raleigh, NC, USA. *e-mail: johan.lindgren@geol.lu.se

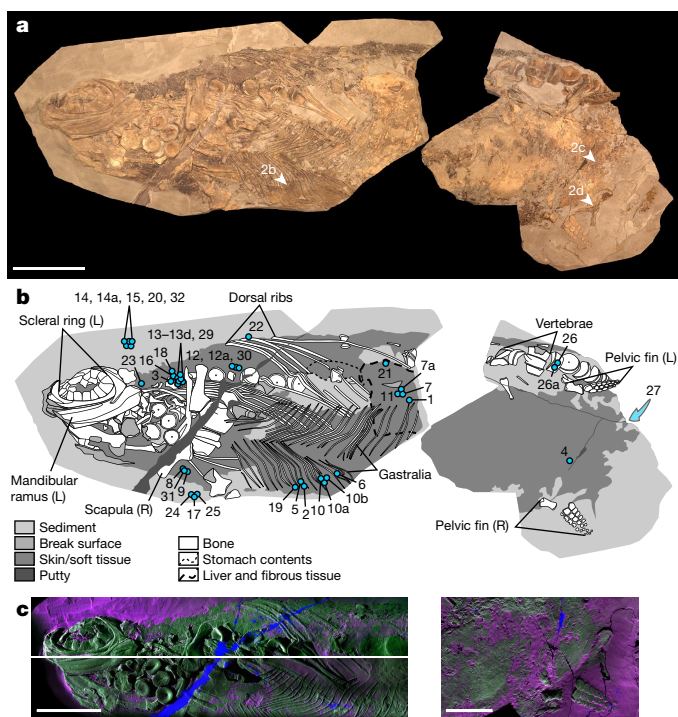


Fig. 1 | *Stenopterygius* specimen MH 432. a, b, Photographic (a) and diagrammatic (b) representation of the fossil in oblique ventral view. Arrowheads indicate enlargements shown in Fig. 2b–d; numerals (blue circles) denote the locations from which samples were taken (sample 27 was collected from the underside of the slab). L, left-hand skeletal element; R, right-hand skeletal element. c, Synchrotron rapid-scanning X-ray fluorescence (SRS-XRF) false-colour images showing the spatial distribution of potassium (magenta), calcium (green) and titanium (blue). Extended Data Fig. 1 presents additional SRS-XRF maps and photographs of MH 432 under regular and ultraviolet light. Scale bars, 5 cm (c right), 10 cm (a, c left).

Cutis and melanophores

The soft tissues of the body outline are preserved as a mineralized, semi-continuous covering (Fig. 1 and Extended Data Fig. 1). No scales or scutes are evident (Fig. 2a–d). Instead, the external surface is smooth, and was presumably comparable in life to the rubbery^{8,9} skin of extant cetaceans (Fig. 2q) and adult individuals of the leatherback sea turtle, *Dermochelys coriacea* (Fig. 2s). Wrinkles and ripples (Fig. 2c, d) closely resemble those that have been observed to distort decomposing integument following loss of structural integrity (Fig. 2r). Demineralization, and subsequent histological and microscopic examination, evinced a multi-layered subsurface architecture (Fig. 2e–p and Extended Data Fig. 2a–e) that corresponds to the laminated epidermis and dermis of modern tetrapods^{10,11} (Fig. 2t). Sparse oval perforations about 70 μm in diameter pass through the cutis (Fig. 2i, j), and are tentatively interpreted as pores.

The approximately 100- μm -thick epidermis retains cell-like structures that are likely to represent preserved melanophores (Fig. 3 and Extended Data Fig. 3) and keratinocytes that were undergoing differentiation^{11,12} (Fig. 2m–o). The outer section of the skin can be further divided into an upper stratum corneum, characterized by stratified squamous cells (Fig. 2m–o and Extended Data Fig. 2b), and a deeper unit that contains polyhedral-to-oblate cellular bodies that become increasingly flattened towards the external surface (Fig. 2o and Extended Data Fig. 2d). This lower part is reasonably interpreted as vestiges of the middle and basal layers of the epidermis, comprising the stratum intermedium and stratum germinativum, respectively; the junction between these layers is not well-defined in MH 432. Vertical extensions at the epidermal–dermal interface form a regularly arranged series of interlocking ridges that run parallel to the longitudinal axis of the body (Fig. 2f and Extended Data Fig. 2a). This microstructure is

consistent with the basement membrane of extant tetrapods⁹, and overlies an approximately 80- μm -thick band of connective tissue fibres that constitutes the superficial dermis (Fig. 2l, p and Extended Data Fig. 2e).

The fossilized pigment cells have long dendritic processes and a granular internal filling (Fig. 3, Extended Data Fig. 3 and Supplementary Video 1). They are virtually identical in size (about 10 μm in diameter, excluding the external projections), geometry and inner fabric to the branched melanophores of extant reptiles (Fig. 3l–q). Our chemical identification of eumelanin (Fig. 3r, s and Extended Data Fig. 3f) further suggests that the granular content represents remnant melanosome organelles (see Supplementary Information).

Preservation of cutis

We posit that the cutis of MH 432 was fossilized through partial replication in authigenic calcium phosphate¹³ (Fig. 1c and Extended Data Figs. 1n, o, 2f, g), an eogenetic mineralization process that enabled the retention of three-dimensional cellular morphologies. However, our experimental removal of the inorganic phase revealed a second preservational mode: in situ molecular transformation (geopolymerization or kero-genization), which is a mechanism whereby labile-to-relatively recalcitrant organic compounds are altered into stable macromolecules by (re-)polymerization, polycondensation and/or defunctionalization reactions^{14–16}. Energy-dispersive X-ray microanalysis (EDX), time-of-flight secondary ion mass spectrometry (ToF-SIMS), nanoscale secondary ion mass spectrometry (NanoSIMS), pyrolysis-gas chromatography with mass spectrometry (Py-GC/MS) and infrared microspectroscopy collectively confirmed the primarily organic composition of the artificially released (and still somewhat flexible; Supplementary Video 2) tissues and cells as being dominated by aliphatic and aromatic hydrocarbon moieties (Fig. 4a, d and Extended Data Figs. 2h, 4a, 5). Such geopolymers are likely to be diagenetic transformation products from a variety of compounds. Although specific precursor molecules are difficult to identify, an endogenous lipid source for at least some of them can reasonably be inferred from the inherent chemical stability of lipid hydrocarbon skeletons, and their demonstrated ability to polymerize in situ; this constitutes a major factor in organic preservation^{14–16}. Our interpretation is corroborated by the apparent preferential preservation of lipid-rich cellular envelopes¹¹ (Fig. 2o and Extended Data Fig. 2d), intracellular vacuoles¹⁷ (Fig. 2k) and extracellular lamellar membranes¹¹ (Fig. 2n and Extended Data Fig. 2b).

Apart from hydrocarbons, traces of the original biomolecular makeup—including cholesterol derivatives (Fig. 4b and Extended Data Fig. 4a) and proteinaceous constituents (Fig. 4a and Extended Data Figs. 5a, b, 6) that retain the immunological characteristics of tropomyosin, haemoglobin, α -keratin, elastin, actin and collagen (Fig. 4e–n and Extended Data Fig. 7a–f)—were detected using a combination of ToF-SIMS, Py-GC/MS, amino acid analysis and in situ immunohistochemistry that incorporated both fluorescence and immunogold labelling. ToF-SIMS and immunohistochemistry further demonstrated co-localization of the geochemical signatures within discrete tissue regions (Fig. 4e–n and Extended Data Figs. 5a, b, 7a–f). Antibody reactivity showed that the various epitopes were spatially distinct, and thus comparable to immunolabelling patterns produced by artificially matured leatherback sea turtle and harbour porpoise (*Phocoena phocoena*) integuments (Extended Data Figs. 7i–t, 9n–q). Although remnant β -keratin has previously been reported from epidermal appendages of other Mesozoic vertebrates¹⁸, we were unable to recognize breakdown products of this durable¹⁹ protein (Extended Data Fig. 7g, h). β -keratin is responsible for the development of hard corneous layers in sauropsid claws, scales and scutes²⁰. Therefore, the apparent absence of this protein in MH 432 supports our microscopy-based observation of a scaleless skin in this ichthyosaur.

Blubber

The subcutaneous layer of MH 432 is over 500 μm thick, and comprises a glossy black material superimposed over a fibrous mat (Fig. 5a).

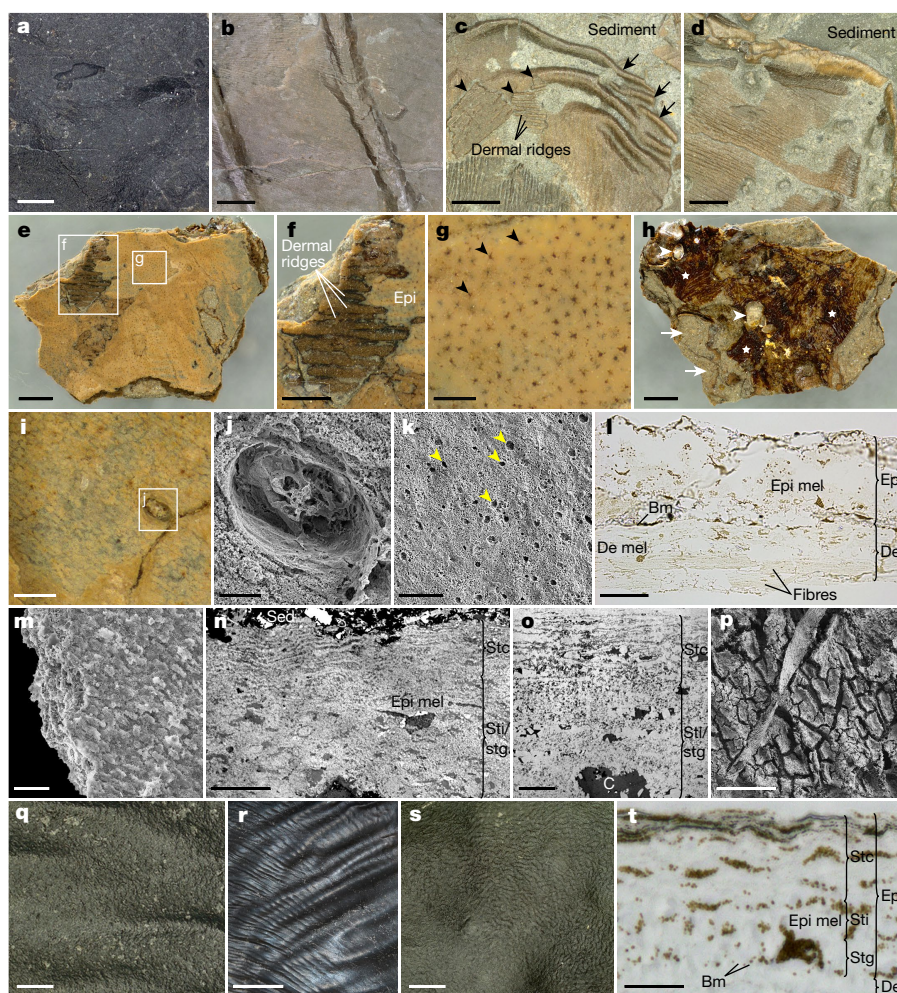


Fig. 2 | Fine structure of integument of *Stenopterygius* specimen MH 432. **a**, External aspect of untreated epidermis (sample 13) showing absence of scales (compare with **q** and **s**). **b**, Skin adhering to the gastralia. **c**, Folding (arrows) and fragmentation (arrowheads) of the integument (compare with **r**). **d**, Trailing edge of the right pelvic fin with post mortem rippling. **e**, Exterior of demineralized flank integument (sample 13a). **f**, Enlargement of interdigitating epidermal and dermal ridges. Epi, epidermis. **g**, Magnified image of subsurface melanophores (arrowheads) embedded in the semi-transparent epidermis. **h**, Internal aspect of demineralized integument showing remnants of blubber and underlying connective tissue (stars). Arrows, infiltrating sediment; arrowheads, authigenic silica minerals. **i**, Demineralized belly skin surface (sample 10) that lacks epidermal melanophores (compare with **g**). **j**, Field emission gun scanning electron microscopy (FEG-SEM) micrograph of a potential pore. **k**, Fractured stratum corneum (sample 13a) with presumed lipid

inclusions preserved as cavities (arrowheads). **l**, Light microscopy section through stratified skin (sample 12a) (compare with **t**). Bm, basement membrane; de, dermis; de mel, dermal melanophore; epi mel, epidermal melanophore. **m**, Oblique transverse section through the stratum corneum (sample 10). **n**, TEM section through pigmented epidermis (sample 13). Sed, sediment; stc, stratum corneum; sti/stg, stratum intermedium and/or stratum germinativum. **o**, TEM section through unpigmented epidermis (sample 10). C, cavity. **p**, FEG-SEM micrograph of dermal fibres in external view (sample 13). **q**, *Delphinus delphis* skin surface. **r**, Post mortem deformation of *D. delphis* skin. **s**, Scaleless carapace skin of adult *D. coriacea*. **t**, Light microscopy section through carapace skin of juvenile *D. coriacea*. Scale bars, 5 μ m (**o**), 20 μ m (**j**, **n**, **t**), 30 μ m (**k**), 50 μ m (**l**, **m**), 100 μ m (**g**, **i**), 200 μ m (**p**), 300 μ m (**f**), 500 μ m (**a**, **e**, **h**, **q**, **s**), 1 mm (**c**, **d**), 2 mm (**b**), 1 cm (**r**).

Histological and chemical analysis showed vertical stratification into a dense, indistinctly laminated carbonaceous unit sandwiched between two sets of surface-parallel seams that are enriched in calcium phosphate (Fig. 5b–f, l, m and Extended Data Fig. 5c). The dark matter is dominated by aliphatic and aromatic compounds (Fig. 5i–n), and appears to represent a highly condensed tissue residue in which geopolymers are so abundant that they obscure virtually all microstructural details when inspected under transmission electron microscopy (TEM) (Fig. 5e, f). This compares well with previously documented fossilized lipid-rich source structures^{14–16}. Moreover, our artificial maturation of harbour porpoise integument (Extended Data Fig. 9e–q) demonstrates that blubber—the peripherally distributed fibro-adipose tissue layer that has only previously been identified in modern cetaceans, sirenians, pinnipeds and the leatherback sea turtle^{21,22}—converts into a similarly condensed mass of flattened cells and fibre bundles when subjected to elevated pressure and

temperature (the epidermis was otherwise relatively unaffected by these experimental procedures; Extended Data Fig. 9f–i). On the basis of the anatomical localization, chemical composition and fabric of the subcutaneous material in MH 432, we interpret it as most probably representing fossilized blubber. Support for this hypothesis includes the presence of potential fatty acid-derived moieties (Extended Data Fig. 4b), as well as the calcium phosphate-reinforced seams (Fig. 5c, d, l, m and Extended Data Fig. 5c) that correspond anatomically to the reticular dermis and superficial fascia of extant balaenid¹⁰ and odontocete⁹ whales. These connective tissue layers consist of closely spaced collagen and elastin strands that enclose the parts of the blubber that are more laden with fat. Although integumental fibrillar proteins do not normally biomineralize²³, they do possess an inherent ability to bind calcium and phosphate ions²⁴, and thus could feasibly have induced the nucleation of calcium phosphate nano-crystallites during the eogenesis of MH 432.

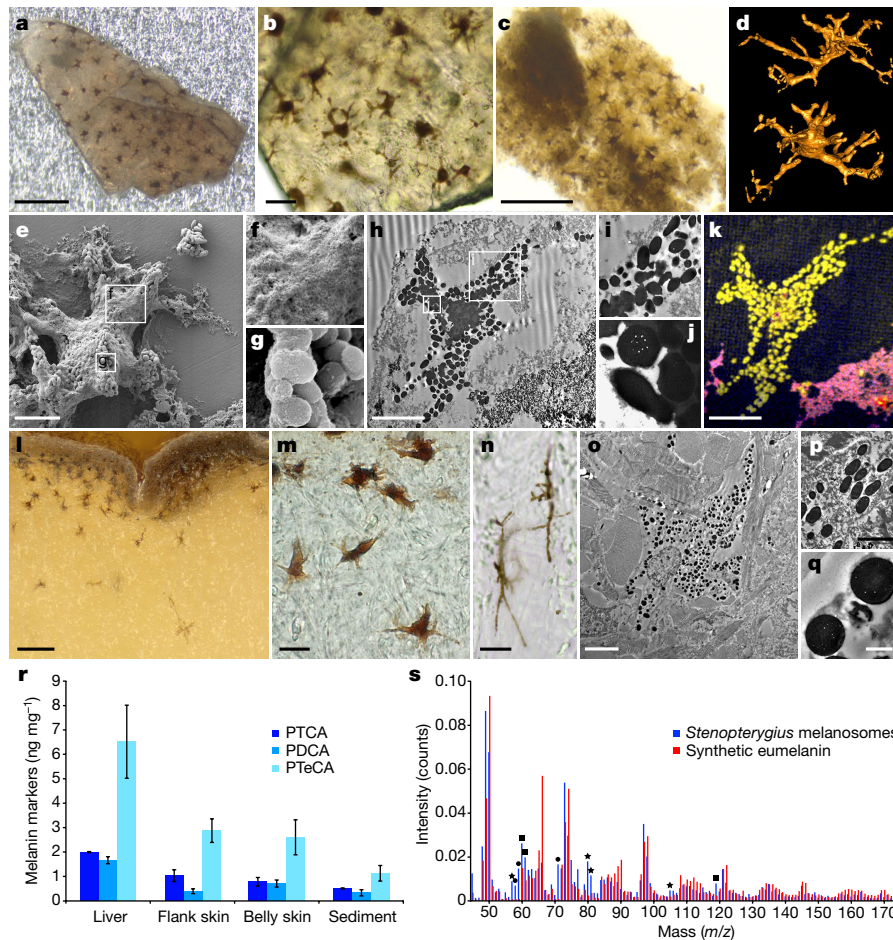


Fig. 3 | Melanophores, melanosomes and eumelanin associated with MH 432. **a**, Light micrograph of epidermal melanophores embedded in semi-transparent phosphatic matrix (sample 13b). **b**, Enlargement of fossil pigment cells and their branching processes (compare with **m**). **c**, Melanophores retained in soft and pliable organic matter after demineralization (sample 13). **d**, Synchrotron-radiation X-ray tomographic microscopy (SRXTM) rendering of a melanophore with dendritic processes (sample 18). **e**, FEG-SEM micrograph of a melanophore liberated from demineralized skin (sample 13). **f**, Magnified outer membrane. **g**, Enlargement showing stacked melanosomes. **h**, TEM micrograph of a branched melanophore that contains melanosome organelles surrounded by matrix (sample 13; compare with **o**). **i**, Magnified cellular extension with dispersed melanosomes (compare with **p**). **j**, Enlargement showing internal melanosome vacuoles (compare with **q**). **k**, False-colour negative ion NanoSIMS image illustrating the

spatial distribution of CH^- (blue), CN^- (yellow) and S^- (red) (see Extended Data Fig. 3e). **l–q**, Melanophores in *D. coriacea* integument imaged under identical conditions as the ichthyosaur pigment cells. Flattened melanophores in **n** straddle a superficial blood vessel. **r**, Quantification of eumelanin markers produced by chemical degradation of MH 432 samples 7a (liver), 13a, 13d (flank skin), 10a, 10b (belly skin) and 14a (sediment) (see Extended Data Fig. 3f). PTCA, pyrrole-2,3,5-tricarboxylic acid; PDCA, pyrrole-2,3-dicarboxylic acid; PTeCA, pyrrole-2,3,4,5-tetracarboxylic acid. **s**, Negative ion ToF-SIMS spectra obtained from sectioned MH 432 melanosomes and synthetic eumelanin (see Supplementary Information). Fossil spectrum peaks that are not consistent with eumelanin represent inorganic ions (squares), epoxy (circles) and sulfur-containing organic ions (stars). Scale bars, 500 nm (**q**), 2 μm (**p**), 5 μm (**e**, **h**, **k**, **o**), 20 μm (**b**, **m**, **n**), 100 μm (**a**, **c**, **l**).

Owing to taphonomic overprinting, the original thickness of the blubber in MH 432 is difficult to estimate. However, on the basis of comparisons with our experimental reduction of a 20-mm-thick blubber layer from a 1.24-m-long harbour porpoise to less than 250 μm under artificially created diagenetic conditions (Extended Data Fig. 9e–j), the depth of the blubber in MH 432 must have been considerable.

Implications

Existing evidence indicates that ichthyosaurs were predominantly dark-coloured in life^{2,3,25}, similar to most air-breathing vertebrates that inhabit pelagic environments today^{8,26,27}. More elaborate patterns involving multiple biochromes and/or photonic nanostructures have been proposed²⁸, but are inconsistent with fossil data^{2,3,25}, as well as the conservative skin palette found in extant oceanic amniotes^{8,27}. Conversely, the conspicuous abundance of melanophores across the flank of MH 432 (Fig. 2g) contrasts with their absence in the belly region (Fig. 2i), which provides compelling evidence that this individual was originally countershaded. A distributional differentiation

of dark and light melanin-based colours between the dorsal and ventral surfaces of the body occurs in many seagoing tetrapods^{8,27}, and functions to provide camouflage²⁹, protect against ultraviolet radiation during sea-surfacing behaviours³⁰ and/or confer thermoregulatory advantages in cold climates³¹. Conflicting reports^{2,25} of monotonal schemes in smaller-bodied individuals of *Stenopterygius* might indicate ontogenetic changes in colour comparable to those seen in living sea turtles²⁶. Moreover, our identification of branched melanophores hints at the possibility that ichthyosaurs were able to physiologically adjust their skin tone via redistribution of melanosomes¹², perhaps to enable ultraviolet filtration, concealment or enhance body temperature stability.

Our morphological and chemical detection of blubber—a hallmark of warm-blooded marine amniotes²¹—provides a rare window into the level of aquatic adaptation achieved by derived ichthyosaurs. In extant animals, blubber accentuates streamlining, buoyancy and serves as a fat store that can be metabolized during periods of energetic stress^{21,22}. However, its primary role as an insulator is crucial to overcome the

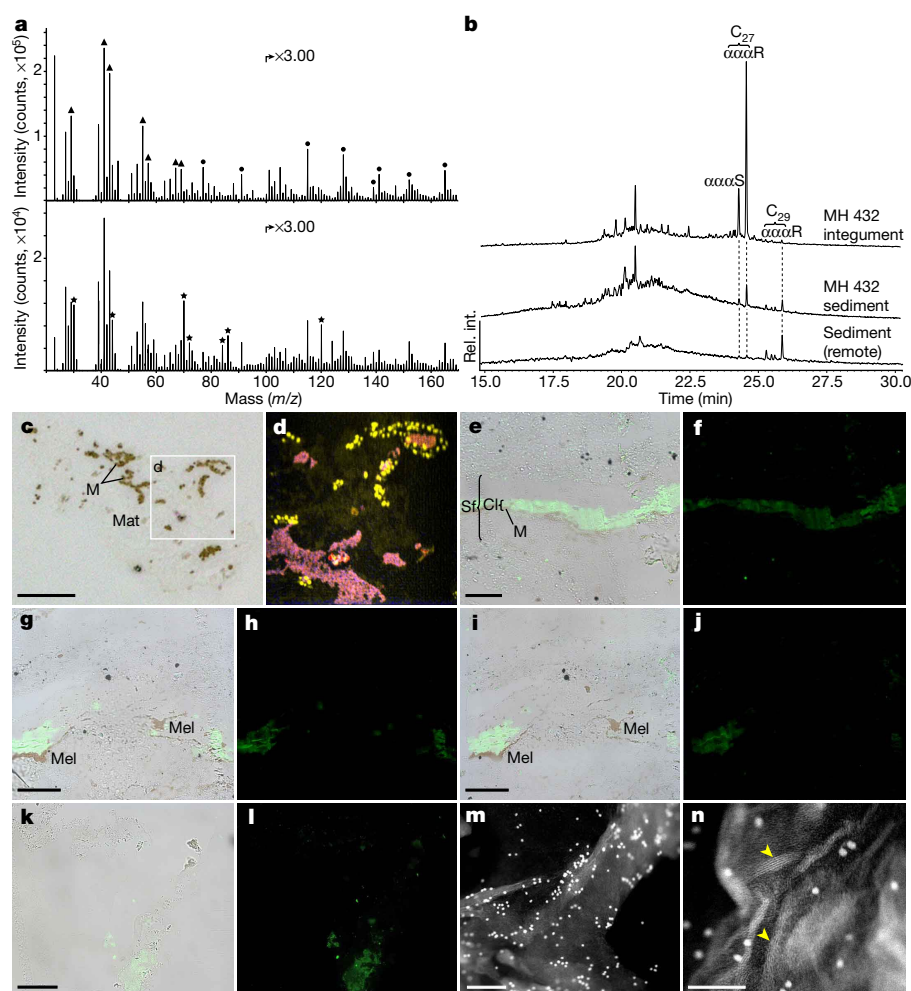


Fig. 4 | Chemistry of the skin of MH 432. **a**, Positive ion ToF-SIMS spectra featuring peaks that are characteristic of aliphatic and polyaromatic hydrocarbons (top spectrum, triangles and circles; sample 2) and of proteinaceous compounds (bottom spectrum, stars; sample 13) (see Extended Data Fig. 5a, b). **b**, Py-GC/MS ion chromatograms (m/z 217, normalized to sample weight) that reveal eukaryote-derived steranes from the integument (sample 16), host rock (sample 14) and sediment sampled at a distance from the fossil (see Extended Data Fig. 4a). Cholestanes (C_{27}), present at high levels in the integument, are interpreted as diagenetic products of ichthyosaur cholesterol. Stigmasteranes (C_{29}), which are present at pronounced levels in the sediment, represent background signal from algae and/or terrestrial plants. $\alpha\alpha\alpha S$ and $\alpha\alpha\alpha R$ represent (20S)-5 α ,14 α ,17 α (H) and (20R)-5 α ,14 α ,17 α (H) isomers, respectively. Rel. int., relative intensity. **c**, **d**, Light microscopy section (**c**) and false-colour negative ion NanoSIMS image (**d**) illustrating the spatial distribution of CH^- (blue), CN^- (yellow) and S^- (red) (see Extended Data

Fig. 2h). M, melanosomes; mat, skin matrix. **e–l**, Immunohistochemical staining of demineralized MH 432 skin material exposed to antisera raised against *Gallus domesticus* tropomyosin (sample 13) (**e**, **f**), *Alligator mississippiensis* haemoglobin (sample 13) (**g**, **h**), *Struthio camelus* haemoglobin (sample 13) (**i**, **j**) and *G. domesticus* α -keratin (sample 8) (**k**, **l**). In **e**, **g**, **i**, **k**, the localization of antibody–antigen complexes is indicated via superimposed green fluorescent signal on transmitted light images of sectioned tissue; in **f**, **h**, **j**, **l**, a fluorescein isothiocyanate filter was used. Both anti-haemoglobin antibodies have a similar binding pattern, and are localized to a discrete structure (possibly a vessel) surrounded by a melanophore (compare with Fig. 3n). Cl, concentration layer; mel, melanophore; sf, skin fold. **m**, **n**, Low-magnification (**m**) and high-magnification (**n**) immunogold labelling of MH 432 skin fibres exposed to anti- α -keratin antibody (sample 12). Arrowheads, filamentous matter (compare with Extended Data Fig. 7s, t). Scale bars, 100 nm (**n**), 200 nm (**m**), 20 μ m (**c**), 30 μ m (**e**, **g**, **i**, **k**).

thermal conductivity of seawater²¹; the presence of this specialized fibro-adipose tissue layer in MH 432 is consistent with reconstructions of ichthyosaurs as homeotherms or regional endotherms^{4,32,33}. Suggestions that these ancient reptiles were capable of mesopelagic deep-diving¹, sustained cruising¹ and toleration of cold-water environments³⁴ likewise accord with the development of blubber as one of many integrated physiological and behavioural mechanisms that are required to maintain a high body temperature that is independent of ambient conditions^{4,25,32,34}.

Our experimental results demonstrate that the integument of *Stenopterygius* had both a smooth external surface and thick subcutaneous layer of fibro-adipose tissue. This is notably similar to modern whales^{9,10} and adult individuals of the leatherback sea turtle⁸, and reveals multiple aspects of soft-tissue convergence that range across a time span of more than 180 million years. In addition, the mosaic of

cetacean and reptilian traits that is characteristic of *Stenopterygius* anatomy recurs in its integumental histology, with the presence of branched melanophores (rather than mammalian melanocytes) and absence of dermal ossifications (otherwise found in the leatherback sea turtle⁸). We attribute these adaptive specializations to the morphological and physiological constraints imposed upon all pelagic tetrapods during their evolutionary transition towards life in the sea.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0775-x>.

Received: 18 March 2018; Accepted: 16 October 2018;
Published online 5 December 2018.

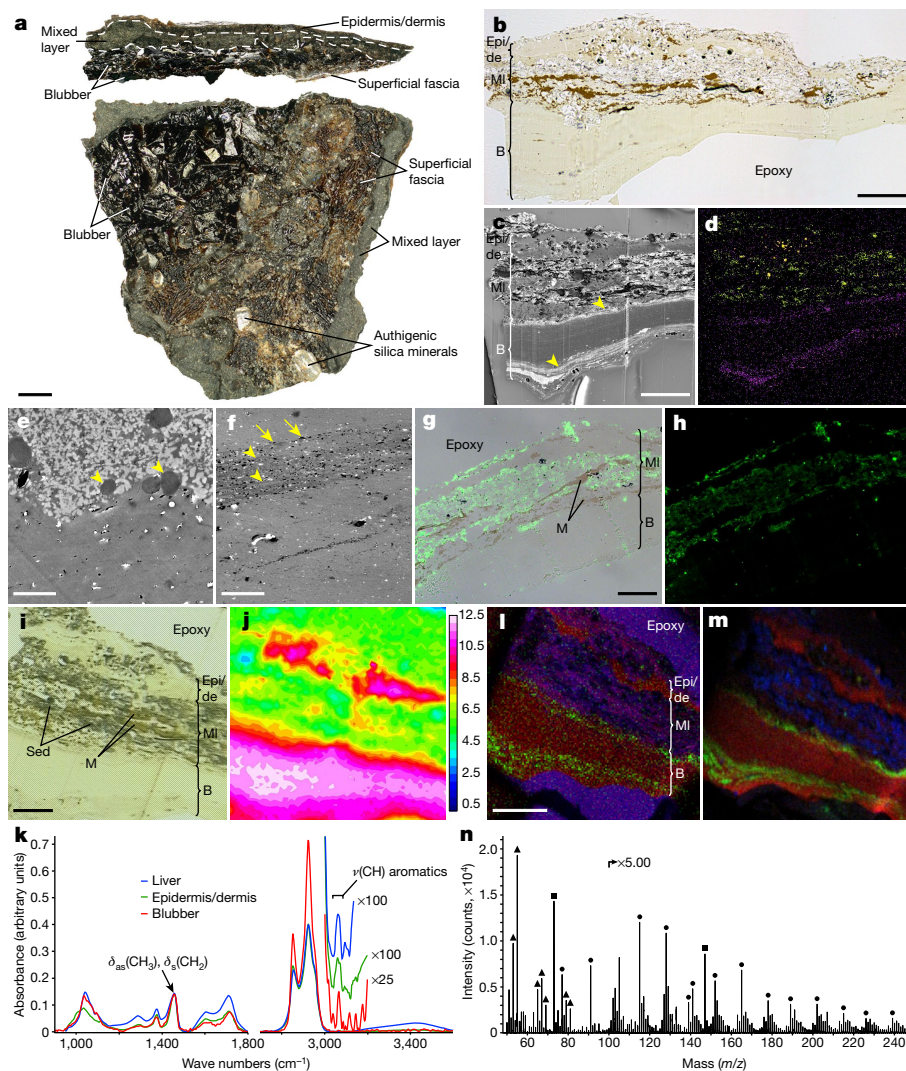


Fig. 5 | Structure and chemistry of MH 432 blubber. **a**, Side (top) and internal (bottom) views of untreated integument (sample 13a). Dashed line demarcates a layer containing both endogenous (skin) and exogenous (clay and bacteria) features (see Extended Data Fig. 10d–h). **b**, Light microscopy section through demineralized integument (sample 13a). B, blubber; epi/de, epidermis and/or dermis; ml, mixed layer (see Supplementary Information). **c**, **d**, Back-scattered electron micrograph (**c**) and three-colour EDX map (**d**) (sample 13a) that indicate the distribution of aluminium (lime), phosphorous (purple) and sulfur (yellow). Arrowheads, reticular dermis (top) and superficial fascia (bottom). **e**, TEM micrograph from the interface between the phosphatized epidermis and/or dermis (top) and polymerized blubber (bottom). Arrowheads, scattered melanosomes. **f**, Organic seams (arrowheads) with calcium phosphate nano-crystallites (arrows) embedded in amorphous blubber residue. **g**, **h**, Immunohistochemical staining for antibodies raised against α -keratin (green fluorescent signal) demonstrating absence of

antibody–antigen complexes in the blubber (sample 13a). M, melanosomes. **i–k**, Infrared absorbance data from the blubber (sample 13a) along with spectra obtained from the epidermis and/or dermis (sample 13a) and liver (sample 1). Band assignments and vibration modes (as, asymmetric; δ , deformation; s, symmetric; ν , stretching) are shown in **k**. Intensities in **j** originate from the 2,805–3,016- cm^{-1} interval denoting $\nu(\text{CH})$ absorption (see Supplementary Information). Sed, sediment (clay). **l**, Positive ion ToF-SIMS image of sample 13a showing the spatial distribution of peaks typical of polyaromatics (red), calcium phosphate (green) and epoxy (blue). **m**, Negative ion ToF-SIMS image with peaks characteristic of sulfur-containing fragments (red), phosphate (green) and silica (blue). **n**, Positive ion ToF-SIMS spectrum recorded from the blubber. Triangles and circles denote peaks typical of aliphatics and polyaromatics, respectively; squares, polydimethylsiloxane (see Supplementary Information). Scale bars, 1 μm (**e**), 5 μm (**f**), 30 μm (**g**, **i**), 50 μm (**b**, **c**, **l**), 500 μm (**a**).

- Motani, R. Evolution of fish-shaped reptiles (Reptilia: Ichthyopterygia) in their physical environments and constraints. *Annu. Rev. Earth Planet. Sci.* **33**, 395–420 (2005).
- Fraas, E. Ueber die Finne von *Ichthyosaurus*. *Jahresh. Vereins vaterl. Naturk. Württemberg* **44**, 280–303 (1888).
- Whitear, M. On the colour of an ichthyosaur. *Ann. Mag. Nat. Hist.* **9**, 742–744 (1956).
- Wiman, C. Über Ichthyosaurier und Wale. *Senckenbergiana* **27**, 1–11 (1946).
- Lingham-Soliar, T. Rare soft tissue preservation showing fibrous structures in an ichthyosaur from the Lower Lias (Jurassic) of England. *Proc. R. Soc. Lond. B* **266**, 2367–2373 (1999).
- Martill, D. M. Prokaryote mats replacing soft tissues in Mesozoic marine reptiles. *Mod. Geol.* **11**, 265–269 (1987).
- Smithwick, F. M., Mayr, G., Saitta, E. T., Benton, M. J. & Vinther, J. On the purported presence of fossilized collagen fibres in an ichthyosaur and a theropod dinosaur. *Palaeontology* **60**, 409–422 (2017).
- Wyneken, J. in *The Leatherback Turtle: Biology and Conservation* (eds Spotila, J. R. & Tomillo, P. S.) 32–48 (John Hopkins Univ. Press, Baltimore, 2015).
- Cozzi, B., Huggenberger, S. & Oelschläger, H. *Anatomy of Dolphins: Insights into Body Structure and Function* (Academic, Amsterdam, 2017).
- Reeb, D., Best, P. B. & Kidson, S. H. Structure of the integument of southern right whales, *Eubalaena australis*. *Anat. Rec.* **290**, 596–613 (2007).
- Landmann, L. in *Biology of the Integument* (eds Bereiter-Hahn, J. et al.) 150–187 (Springer, Berlin, 1986).
- Alibardi, L. Ultrastructural features of skin pigmentation in the lizard *Heloderma suspectum* with emphasis on xanto-melanophores. *Acta Zool.* **96**, 154–159 (2015).
- Wilby, P. R. & Briggs, D. E. G. Taxonomic trends in the resolution of detail preserved in fossil phosphatized soft tissues. *Geobios* **30**, 493–502 (1997).
- Gupta, N. S., Tetlie, O. E., Briggs, D. E. G. & Pancost, R. D. The fossilization of eurypterids: a result of molecular transformation. *Palaios* **22**, 439–447 (2007).

15. Gupta, N. S., Cody, G. D., Tetlie, O. E., Briggs, D. E. G. & Summons, R. E. Rapid incorporation of lipids into macromolecules during experimental decay of invertebrates: initiation of geopolymer formation. *Org. Geochem.* **40**, 589–594 (2009).
16. O'Reilly, S., Summons, R., Mayr, G. & Vinther, J. Preservation of uropygial gland lipids in a 48-million-year-old bird. *Proc. R. Soc. Lond. B* **284**, 20171050 (2017).
17. Jackson, M. K. & Sharawy, M. Lipids and cholesterol clefts in the lacunar cells of snake skin. *Anat. Rec.* **190**, 41–45 (1978).
18. Pan, Y. et al. Molecular evidence of keratin and melanosomes in feathers of the Early Cretaceous bird *Eoconfuciusornis*. *Proc. Natl Acad. Sci. USA* **113**, E7900–E7907 (2016).
19. Moyer, A. E., Zheng, W. & Schweitzer, M. H. Keratin durability has implications for the fossil record: results from a 10 year feather degradation experiment. *PLoS ONE* **11**, e0157699 (2016).
20. Toni, M. & Alibardi, L. Soft epidermis of a scaleless snake lacks beta-keratin. *Eur. J. Histochem.* **51**, 145–151 (2007).
21. Iverson, S. J. in *Encyclopedia of Marine Mammals* (eds Perrin, W. F. et al.) 115–120 (Academic, Amsterdam, 2009).
22. Davenport, J. et al. Fat head: an analysis of head and neck insulation in the leatherback turtle (*Dermochelys coriacea*). *J. Exp. Biol.* **212**, 2753–2759 (2009).
23. Fartasch, M., Haneke, E. & Hornstein, O. P. Mineralization of collagen and elastic fibers in superficial dystrophic cutaneous calcification: an ultrastructural study. *Dermatologica* **181**, 187–192 (1990).
24. Balakrishnan, S. et al. Studies on calcification efficacy of stingray fish skin collagen for possible use as scaffold for bone regeneration. *Tissue Eng. Regen. Med.* **12**, 98–106 (2015).
25. Lindgren, J. et al. Skin pigmentation provides evidence of convergent melanism in extinct marine reptiles. *Nature* **506**, 484–488 (2014).
26. Wyneken, J. *The Anatomy of Sea Turtles* (U.S. Department of Commerce NOAA Technical Memorandum NMFS-SEFSC-470, 2001).
27. Shriahai, H. & Jarrett, B. *Whales, Dolphins, and Other Marine Mammals of the World* (Princeton Univ. Press, Princeton, 2006).
28. Vinther, J. A guide to the field of palaeo colour: melanin and other pigments can fossilise: reconstructing colour patterns from ancient organisms can give new insights to ecology and behaviour. *BioEssays* **37**, 643–656 (2015).
29. Marshall, J. & Johnsen, S. in *Animal Camouflage: Mechanisms and Function* (eds Stevens, M. & Merilaita, S.) 186–211 (Cambridge Univ. Press, Cambridge, 2011).
30. Martinez-Levasseur, L. M. et al. Acute sun damage and photoprotective responses in whales. *Proc. R. Soc. Lond. B* **278**, 1581–1586 (2011).
31. James, M. C., Myers, R. A. & Ottensmeyer, C. A. Behaviour of leatherback sea turtles, *Dermochelys coriacea*, during the migratory cycle. *Proc. R. Soc. Lond. B* **272**, 1547–1555 (2005).
32. Bernard, A. et al. Regulation of body temperature by some Mesozoic marine reptiles. *Science* **328**, 1379–1382 (2010).
33. Nakajima, Y., Houssaye, A. & Endo, H. Osteohistology of the Early Triassic ichthyopterygian reptile *Utatusaurus hataii*: implications for early ichthyosaur biology. *Acta Palaeontol. Pol.* **59**, 343–352 (2014).
34. Kear, B. P. Marine reptiles from the Lower Cretaceous of South Australia: elements of a high-latitude cold-water assemblage. *Palaeontology* **49**, 837–856 (2006).

Acknowledgements We thank E. R. Schroeter for scientific advice. S. M. Webb and C. Roach assisted during the SRS-XRF analysis. T. Wigren produced the artistic reconstructions. Financial support was provided by a Grant for Distinguished Young Researchers (642-2014-3773, Swedish Research Council) to J.L., as well as a National Science Foundation INSPIRE grant (EAR-1344198) to M.H.S. and W.Z., and donations from F. M. and S. P. Orr, and V. and G. Mullis. The Paul Scherrer Institute, Switzerland, provided beamtime at the TOMCAT beamline of the Swiss Light Source. SRS-XRF data were collected at the Stanford Synchrotron Radiation Lightsource using beamline 10-2. NanoSIMS measurements were obtained at the Chemical Imaging Infrastructure, Chalmers University of Technology and University of Gothenburg, which is supported by the Knut and Alice Wallenberg Foundation. Part of this work was performed at the Analytical Instrumentation Facility (AIF) of North Carolina State University. The AIF is supported by the State of North Carolina and the National Science Foundation (ECCS-1542015) and is a member of the North Carolina Research Triangle Nanotechnology Network, a site within the National Nanotechnology Coordinated Infrastructure.

Reviewer information *Nature* thanks A. Houssaye, B. Kessler, S. Kiel and R. N. S. Sodhi for their contribution to the peer review of this work.

Author contributions J.L. conceived the project. The text was written by J.L. and B.P.K. with contributions from S.S., P.S., M.H.S. and P.E.A., and feedback from all authors. All figures were assembled by J.L. with input from all authors. R.H. collected and prepared MH 432. J.L., P.S. and M.J. recorded the SRS-XRF measurements; O.G., P.S., C.A., T.K. and J.L. conducted the FEG-SEM and TEM analyses; V.T. performed the Py-GC/MS experiments; S.I. and K.W. carried out the alkaline hydrogen peroxide oxidation analysis; M.H.S. and W.Z. conducted the immunohistochemistry investigation; P.S. and J.L. performed the ToF-SIMS experiments; A.T., P.M., J.L. and P.S. carried out the NanoSIMS analysis; A.E., J.L. and P.U. recorded the infrared microspectroscopic measurements; M.E.E. and F.M. acquired the SRXTM data; M.O. performed the amino acid analysis; and M.J. and I.R.-M. conducted the maturation experiments.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0775-x>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0775-x>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to J.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Fossil material. Before analysis, MH 432 was rinsed multiple times with 96% ethanol and Milli-Q water to remove potential contaminants from human handling. Tissue and sediment samples were collected from the fossil using sterile instruments and washed successively in 96% ethanol and Milli-Q water. They were subsequently dried under a hood, wrapped loosely in aluminium foil, and stored in sealed, sterile containers. Samples selected for light microscopy, FEG-SEM, TEM, NanoSIMS, ToF-SIMS and infrared microspectroscopy were demineralized using EDTA (0.5 M, pH 8.0) for about four weeks with regular buffer changes to dissolve the mineral phase. Debris from the demineralization process was rinsed multiple times with Milli-Q water to discard remaining EDTA. Fresh aluminium foil was used to cover all work areas, and surgical gloves were worn during all handling and treatment. Experimental parameters between ancient and modern (see below) materials were identical; however, both sets of samples were treated in segregated laboratory spaces.

Modern reference materials. Frozen tissue samples from a sub-adult male harbour porpoise, *P. phocoena* (A2016/05526), found dead in Swedish waters in 2016, were provided by the Department of Environmental Research and Monitoring, Environmental Specimen Bank, Swedish Museum of Natural History, Stockholm. Similarly, integumental samples from a juvenile male short-beaked common dolphin, *D. delphis* (SAM M26880), washed up dead in South Australia in 2012, were obtained from the South Australian Museum (SAM), Adelaide. A juvenile leatherback sea turtle, *D. coriacea* (ZMUC-R2106), stored in 70% ethanol, was acquired as a gift from Danmarks Akvarium in 1962, and has since been part of the vertebrate collection at the Zoological Museum, Natural History Museum of Denmark (ZMUC), Copenhagen. The beached carcass of a second (adult female) *D. coriacea* (ZMUC-R2112) was recovered in Denmark in 1979. The integument from ZMUC-R2112 was cleaned and dried, and is currently housed in the collections at the ZMUC.

Maturation experiments. Integument samples from the juvenile leatherback sea turtle (Extended Data Fig. 9a) and harbour porpoise (Extended Data Fig. 9e) were placed in cylindrical borosilicate glass vials with the epidermis facing downwards. The samples were then pressed from the hypodermis side using glass rods with flattened ends that were inserted into the vials. Each vial was placed in a larger beaker, and another smaller beaker (containing 600 g sand) was placed on top of the glass rod to compact the integument. The whole apparatus was placed in a ventilated oven heated to 80 °C. The samples were inspected at regular intervals (after 4 h on the first day, and then once on each following day), and the colourless oil expelled from the porpoise blubber was removed using a glass pipette. The samples were heated in this same manner for six consecutive days, during which the fatty inner tissues were compressed into a brittle amber-coloured substance (the epidermis was otherwise only slightly transformed by this same treatment; Extended Data Fig. 9f, g).

To artificially accelerate the maturation process, we also exposed the samples to higher temperatures and pressures. Each compacted sample was first cut into four smaller sections, with one wrapped in aluminium foil and then placed in an autoclave. The autoclave consisted of a stainless-steel high-pressure vessel (30 × 10-mm column, Applied Porous Technologies) connected to an argon gas cylinder. The autoclave was purged three times by filling it to 80 bar with argon gas, before being relieved through the exhaust pipe. The autoclave was subsequently filled to approximately 75% of the desired final pressure, and heated in a thermostat-controlled oven (GC oven, 5890 series II, Hewlett Packard). After a few minutes (when the system had reached the desired temperature), a manual valve was used to adjust the final pressure to 200 bar. The samples were first exposed to 100 °C and 200 bar for 24 h, and then to 150 °C and 200 bar for 96 h; this second exposure was necessary because no further morphological changes were apparent after treatment at 100 °C and 200 bar. These procedures were found to darken the porpoise blubber (Extended Data Fig. 9h). Sample sections exposed to 200 °C and 200 bar for 24 h either disintegrated completely or turned into a sticky semi-solid substance, which was not examined further.

Light microscopy. Tissue samples from our comparative selection of extant animals were inundated in freshly prepared fixative solution, 2% paraformaldehyde and 2.5% glutaraldehyde in 0.1 M cacodylate buffer (pH 7.4) for 24 h at 4 °C. The samples were then dehydrated in a graded ethanol series and embedded in epoxy resin (Agar 100, Resin kit R1031) via acetone, which was left to polymerize for 48 h at 60 °C. The demineralized fossil tissue was immersed in epoxy resin (AGAR 100, resin kit R1031), which was left to polymerize at room temperature for 72 h, followed by 48 h at 60 °C. Semi-thin (1.5-µm) light-microscopic sections were then cut with a glass knife using a Leica EM UC7 Ultramicrotome, and mounted on objective glasses. Every second section was stained with Richardson's solution, before examination using an Olympus CX21 microscope. Micrographs were recorded using a Lumenera Infinity 2-IRC CCD camera with Lumenera Infinity Analyze version 6.4.1 software.

FEG-SEM and EDX. Untreated and demineralized fossil samples were mounted on conductive tape, sterile silicon wafers or CaF₂ infrared windows. Following

ToF-SIMS and/or infrared microspectroscopic measurements (see below), the samples were coated with gold or platinum/palladium using a Cressington 208HR High Resolution Sputter Coater. They were then studied in a Tescan Mira3 High Resolution Schottky FEG-SEM with both standard and in-lens secondary electron and back-scattered electron detectors at acceleration voltages between 1 and 15 kV, and a working distance of 3–10 mm. Elemental analyses and mappings were performed with an energy-dispersive spectrometer (X-MaxN 80, 124 eV, 80 mm²) from Oxford Instruments, linked to the instrument. Additional investigations were conducted on gold/palladium-coated samples in a Zeiss Supra 40VP FEG-SEM (2 keV electron energy, 3–5 mm working distance, Everhart–Thornley secondary electron detector).

TEM. Tissue samples from our comparative selection of extant animals were fixed and embedded in epoxy resin as described in 'Light microscopy'. Ultra-thin (50-nm) sections were cut using a Leica EM UC7 Ultramicrotome equipped with a diamond knife, and mounted on pioloform-coated copper grids without further treatment or staining. Demineralized fossil tissue was likewise embedded in epoxy resin (see 'Light microscopy'). Ultra-thin (50-nm) sections were cut and mounted on copper grids without further treatment or staining. All sections were examined in a JEOL JEM-1400 PLUS TEM at 100 kV. Micrographs were recorded with a JEOL Matatoki CMOS camera using TEM Centre for JEM1400 Plus software.

In situ immunohistochemistry. Extant animal and demineralized fossil tissues were embedded in LR White (hard grade, EMS, Cat# 14383) according to the manufacturer's specifications. Sections (200 nm thick) were taken on a Leica EM UC6 Ultramicrotome for in situ immunohistochemistry, whereas 90-nm sections were used for post-embedding TEM immunogold labelling. The 200-nm sections were transferred to 6-well Teflon-coated slides and dried before being etched with 25 µg/ml proteinase K and 0.5 M EDTA for epitope retrieval. This was followed by incubation in 1 mg/ml sodium borohydride to quench autofluorescence. Non-specific binding was prevented by incubating sections in 4% normal serum before immune labelling with each primary antibody overnight at 4 °C; working dilutions followed specifications recommended by the manufacturers. After repeated washing to remove unbound antibodies, the sections were incubated with a biotin-conjugated secondary antibody for 2 h at room temperature. An avidin-biotin reagent (Vector Laboratories A-2001) was applied for 1 h. All incubations were separated by sequential washes using 1 × PBS. Finally, all sections were mounted with anti-fade mounting medium (Vector H-1000), and coverslips were applied. Examinations were undertaken using a Zeiss Axioskop 2 Plus biological microscope.

Immunogold assays were performed at room temperature on both the extant and fossil tissue samples. The fossil tissues were prepared in a specifically designed laboratory in which no modern animal tissues are permitted; conversely, the extant-animal samples were prepared in a separate laboratory. Sections of 90-nm thickness were collected on carbon-coated nickel grids (EMS, Cat# CFT200-NI). The grids were then incubated in: (1) PBS buffer with 1% Tween20 for 10 min; (2) 4% donkey serum in PBS (1 h); (3) primary antibodies against α-keratin (1:10) for 3 h; (4) washed with PBS buffer with 1% Tween20 10 times; and (5) 18 nm Colloidal Gold AffiniPure donkey anti-goat IgG (H + L) 1:10 (Jackson Immuno Research, Cat# 705-215-147). After incubation with primary antibodies, the grids were washed using the procedures outlined above, and then stained in methanolic uranyl acetate followed by Reynold's lead citrate, and inspected using an aberration-corrected STEM-FEI Titan 80-300 electron microscope.

SRXTM. SRXTM was performed at the TOMCAT beamline of the Swiss Light Source, Paul Scherrer Institute, Switzerland. Samples were mounted on either low-light-refractive, 0.3-mm fishing lines (Berkley, Trilene super strong, 100% fluorocarbon) or wooden toothpicks. The beam energy was set to 12 or 20 keV, depending on sample size. The transmitted X-rays were converted into visible light by a 5.9-µm-thick Tb-doped LSO scintillator screen (FEE). Projection images (1,200 over 180°) were magnified by microscope optics (40×), and digitized by a sCMOS camera (PCO.edge 5.5) with a 2,560 × 2,160 pixel chip, a 6.5-µm pitch, and a 16-bit dynamic range. Tomographic reconstructions followed single-distance phase retrieval of the acquired projections, and were performed using a Fourier transform method routine, and a gridding procedure. The resulting tomographic volumes had an isotropic voxel size of 0.16 µm, and were rendered using Voxler 2 and 3.

SRS-XRF. X-ray fluorescence images were collected at the Stanford Synchrotron Radiation Lightsource using beamline 10-2. The incident X-ray energy was set to 11.0 keV using a Si (111) double-crystal monochromator with the Stanford Positron Electron Accelerating Ring storage ring containing 500 mA at 3.0 GeV. The fluorescence lines of the targeted elements, as well as the intensity of the total scattered X-rays, were monitored using a silicon drift Vortex detector (SII NanoTechnology). The focused beam of 100 × 100 µm was determined by a 100-µm Ta pinhole aperture. The incident and transmitted X-ray intensities were measured with nitrogen-filled ion chambers, and the incident X-ray flux on the sample was measured at ~8 × 10¹⁰ ph/s. The samples were mounted at 45° relative to the incident X-ray beam, with the fluorescence detector mounted at 90° relative to

the incident beam. The sample and detector were maintained within an ambient atmospheric environment, with the detector held at a distance of ~ 4 cm from the sample. The samples were mounted vertically on an aluminium bracket, and spatially rastered in the X-ray beam while data were collected continuously during stage motion. The fluorescence signal was gated by the encoder signal at pixel sizes of 100 or 200 μm . Beam exposure time was 20 ms per pixel.

Infrared microspectroscopy. Infrared microspectroscopic measurements were recorded at the Department of Biology, Lund University. A Hyperion 3000 microscope combined with a Tensor 27 spectrometer was used, together with either a single element mercury cadmium telluride (MCT) detector ($250 \times 250 \mu\text{m}$), or a 64×64 pixel focal plane array (FPA) detector and a Global light source. The microscope was operated in transmission mode at 4 cm^{-1} resolution, and a $15\times$ objective was used. One hundred and twenty-eight individual scans were averaged to achieve a good signal-to-noise ratio. Sample 13 (which includes the isolated melanophore depicted in Fig. 3e–g and Extended Data Fig. 3d, g) was also analysed at the SMIS beamline, Synchrotron SOLEIL, France. This facility uses an Agilent Cary 620 FTIR microscope system with a 128×128 FPA MCT detector (4-cm^{-1} resolution, $25\times$ objective).

NanoSIMS. Ion images were acquired using the Cameca NanoSIMS 50L instrument at the Chemical Imaging Infrastructure, Chalmers University of Technology and University of Gothenburg. Measurements were performed with a 16 keV Cs^+ primary ion beam, rastering the sample surface using a 1.5 pA (aperture diaphragm D1 2) primary current. The spatial resolution of the primary beam size was 200 nm. The count rates of the $^{12}\text{C}^+\text{H}^-$, $^{12}\text{C}^{14}\text{N}^-$, $^{28}\text{Si}^-$, $^{31}\text{P}^-$, $^{32}\text{S}^-$ and $^{40}\text{Ca}^{16}\text{O}^-$ were measured simultaneously in multicollection mode using electron multipliers. Mass-filtered images were then acquired using entrance slit 3 (width, 20 μm) and aperture slit 2 (width, 200 μm); the energy slit was kept fully open (energy band pass, ≤ 100 eV). The relative transmission of the mass spectrometer is $\sim 38\%$ with a mass resolving power of 9,000 on $^{28}\text{Si}^-$ (CAMECA definition). Each image is composed of 10 planes (for each measured mass). Images were processed using ImageJ and the OpenMIMS plugin. The count rates were corrected using a dead time of 44 ns for each electron multiplier.

ToF-SIMS. ToF-SIMS analyses were conducted in the static SIMS mode on a TOFSIMS IV instrument (IONTOF GmbH) using 25 keV Bi_3^+ primary ions and low-energy electron flooding for charge compensation. Positive and negative ion data were acquired with the instrument optimized for either high mass-resolution ($m/\Delta m \sim 5,000$, spatial resolution $\sim 3\text{--}4 \mu\text{m}$) or high image-resolution ($m/\Delta m \sim 300$, spatial resolution $\sim 0.2\text{--}0.5 \mu\text{m}$), because these properties cannot be optimized simultaneously without inordinate increases in analysis time³⁵. The pulsed primary ion current was set at 0.10 pA for the high mass-resolution data, and 0.04 pA for the high image-resolution data.

Py-GC/MS. Untreated fossil and sediment samples were ground to powder and loaded on a fast-heating Pt-filament of a Pyrola 2000 device (Pyrolab SB) coupled to a Varian CP3800 GC and a Varian 1200L MS. Two GC/MS runs were conducted on each sample. First, the filament was heated in 2 ms from 200 to 330 °C (hold 40 s), and a GC/MS run was performed to record the resulting volatile bitumen compounds and contaminants (if present). The sample remained in the pyrolysis device throughout the duration of this first run. In the second run, the temperature was increased to 560 °C (hold 10 s) within 2 ms to release compounds and moieties that were tightly bound to the organic matter (runs shown in Fig. 4b and Extended Data Fig. 4). Samples were weighed onto the pyrolysis filament (~ 5 mg for each run). As a monitor of system performance, 40 ng of perdeuterated *n*-icosane ($\text{C}_{20}\text{D}_{42}$; Sigma) dissolved in 4 μl *n*-hexane was added to each sample. Here, we show pyrograms that exhibited high transfer efficiency consistent with the standard (Fig. 4b and Extended Data Fig. 4). Chromatograms were normalized to the sample weight loaded onto the pyrolysis filament (Fig. 4b and Extended Data Fig. 4b), or adjusted to the highest *n*-alkane peak to illustrate the abundance of the key compound, (20R)-5 α ,14 α ,17 α (H)-cholestane, relative to other pyrolysis products (Extended Data Fig. 4a). This molecule was identified by comparison against, and co-injection of, a commercially available standard (Chiron Laboratories AS). Identification of the 20S isomer was based on its retention time and mass spectral characteristics; minor co-elution with other isomers may have occurred. The gas chromatograph was equipped with a Phenomenex Zebtron ZB-5 capillary column (30 m, 0.1 μm film thickness, inner diameter 0.25 mm), and used helium (1.7 ml/min) as the carrier gas. Pyrolysis products were flushed onto the GC column at an injector temperature of 300 °C, and a split rate of 20. The GC oven temperature was ramped from 40 (3 min) to 310 °C at $10^\circ\text{C min}^{-1}$, and was held for 15 min. Electron-ionization mass spectra were recorded at 70 eV in full-scan mode (mass range 50–450, scan time 0.35 s).

Alkaline hydrogen peroxide oxidation. Untreated fossil and sediment samples were mechanically ground into a powder. Approximately 2–6 mg of each sample was heated with 500 μl 6 M HCl at 110 °C for 16 h in a 10-ml test tube. Following addition of 1 ml water, the tube was centrifuged at 14,000g for 10 min, and the residue then washed with 1 ml water and subjected to alkaline hydrogen peroxide ox-

idation³⁶. One hundred microlitres of water, 375 μl 1 M K_2CO_3 and 25 μl 30% H_2O_2 were added to each tube. Following vigorous mixing at 25 °C for 20 h, the residual H_2O_2 was decomposed by adding 50 μl 10% Na_2SO_3 , and the resulting mixture acidified with 140 μl 6 M HCl. The oxidation mixture was centrifuged at 10,000g for 1 min, and an aliquot (80 μl) of the supernatant was injected directly into a high-performance liquid chromatography (HPLC) system. The setup included a JASCO 880-PU liquid chromatograph (JASCO), a Shiseido C18 column (Capcell Pak, Type MG; 4.6×250 mm; 5- μm particle size; Shiseido), and a JASCO UV detector, which was monitored at 269 nm. The mobile phases comprised 0.1 M potassium phosphate buffer (pH 2.1): methanol, 99:1, at 45 °C for PTCA and PDCA and 0.4 M HCOOH : methanol, 85:15, at 35 °C for PTeCA.

Amino acid analysis. Powdered samples ($\sim 6\text{--}10$ mg) were placed in vacuum hydrolysis test tubes (Kontes Glass, #896860-8910), and 1 ml 6 M HCl containing 1% phenol was added to each tube. The tubes were then evacuated, sealed and placed in an oil bath at 110 °C for 24 h. The hydrolysates were extracted twice with 2 ml ether, and evaporated to dryness under vacuum. The residues were then dissolved in 200–400 μl 0.1 M HCl, and centrifuged at 10,000g. Aliquots (4–5 μl) were applied to LC/MS for quantitative amino acid analysis. Two solvent systems (condition I and II) were used because of variable amino acid sensitivity: condition I for alanine, glutamic acid, glycine and serine; and condition II for all other amino acids. Standard amino acid solutions of 10 and 50 μM in 0.1 M HCl were used as references. Liquid chromatographic conditions included an Agilent 1200 HPLC system (Hewlett Packard) with an Intrada Amino Acid ($2 \times 75\text{-mm}$) (Imtakt) column, and a sequence of solvents: condition I, A = MeCN:THF:25 mM HCOONH_4 : HCOOH (9:75:16:0.3), B = MeCN:100 mM HCOONH_4 (20:80), 0% B (0–7 min), 0–17% B (7–21 min), 100% B (21–35 min); condition II, A = MeCN: HCOOH (100:0.1), B = 100 mM HCOONH_4 , 15% B (0–3 min), 15–100% B (3–20 min), 100% B (20–24 min). The flow rate was set at 0.2 ml/min. Mass spectrometry conditions incorporated a HCTplus (Bruker Daltonics) with an ESI positive source, an ionization nebulizer at 30 psi, dry gas at 7.0 l/min, dry temperature at 320 °C and an ion-trap analyser.

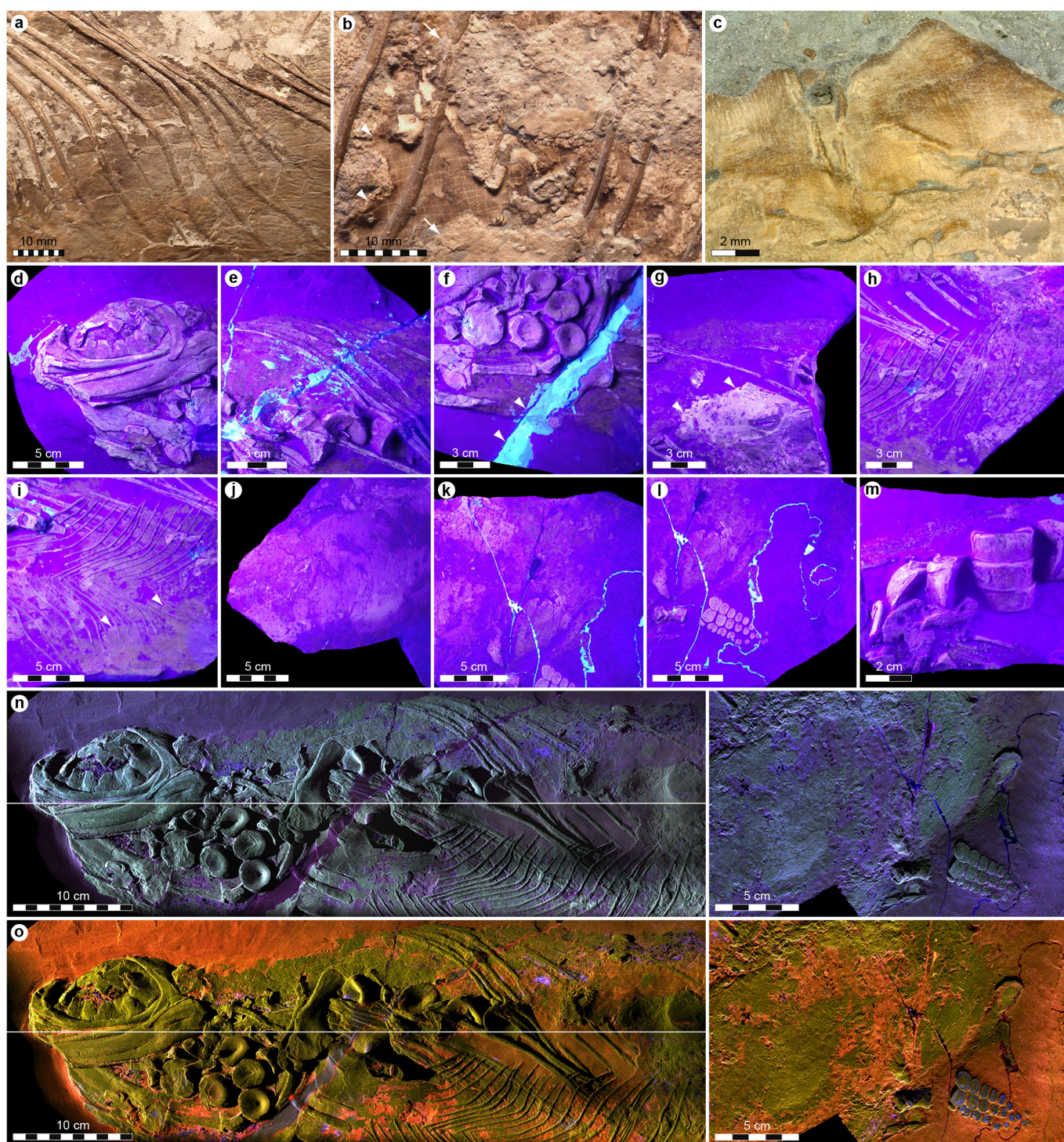
Statistics and reproducibility. Forty-one samples were collected from MH 432 and its associated matrix (Fig. 1b); 18 of these were demineralized. All samples were photographically documented (Figs. 2a, e–i, 3a–c, 5a and Extended Data Fig. 2a), together with comparative tissues from extant vertebrates (Fig. 2q–s and Extended Data Fig. 9e). The following experiments were repeated independently with similar results. Light microscopy (histological sections): MH 432 (Figs. 2l, 4c, 5b, i and Extended Data Figs. 8b, 10d) nine samples; *Dermochelys* (Figs. 2t, 3l–n and Extended Data Fig. 9a–d) nine samples; and *Phocoena* (Extended Data Fig. 9g, i, j) six samples. FEG-SEM: MH 432 (Figs. 2j, k, m, p, 3e–g, 5c and Extended Data Figs. 3b, c, 8c, d) eight samples. EDX: MH 432 (Fig. 5d and Extended Data Figs. 2f, g, 3d) eight samples. TEM: MH 432 (Figs. 2n, o, 3h–j, 5e, f and Extended Data Figs. 2b–e, 8i, j, 10e–h, j) nine samples; *Dermochelys* (Fig. 3o–q) nine samples; and *Phocoena* (Extended Data Fig. 9k–m) six samples. Immunohistochemistry in Fig. 4e–n: for MH 432 sample 13, tropomyosin four times, ostrich haemoglobin four times and alligator haemoglobin three times; for MH 432 sample 8, α -keratin four times; and for MH 432 sample 12, immunogold α -keratin three times. Immunohistochemistry in Fig. 5g, h: MH 432 sample 13a, α -keratin two times. Immunohistochemistry in Extended Data Fig. 7: MH 432 samples 8 and 12a, elastin two times, actin three times, collagen two times, and β -keratin two times; *Dermochelys*, tropomyosin two times, collagen two times, alligator haemoglobin two times, ostrich haemoglobin two times, α -keratin four times and immunogold α -keratin three times. Immunohistochemistry in Extended Data Fig. 8e–h: MH 432 sample 7, ostrich haemoglobin four times and alligator haemoglobin three times. SRXTM in Fig. 3d and Extended Data Fig. 3a, five times. Infrared microspectroscopy: the intensity map (Fig. 5j) was generated from 128 individual scans; each spectrum in Fig. 5k represents an average of three spots; and the melanophore spectrum (Extended Data Fig. 3g) is an average of 32 individual scans (repeated twice). NanoSIMS: melanophore (Fig. 3k and Extended Data Fig. 3e) six times; light microscopy section (Fig. 4d and Extended Data Fig. 2h) four times. ToF-SIMS: eumelanin identification (Fig. 3s) four times; hydrocarbon and peptide/protein identification (Fig. 4a and Extended Data Fig. 5a, b) nine times; blubber (Fig. 5l–n and Extended Data Fig. 5c) four (positive ions) and six (negative ions) times; and liver (Extended Data Fig. 8k) six times. Py-GC/MS: three (integument, liver and sediment) and four (remote sediment) times (Fig. 4b and Extended Data Fig. 4). Alkaline hydrogen peroxide oxidation: two (liver, belly skin and sediment) and three (flank skin) times (Fig. 3r and Extended Data Fig. 3f; data in Fig. 3r are centre values with standard error of mean). Amino acid analysis: two times (Extended Data Fig. 6 and Supplementary Table 2). The other experiments—including the SRS-XRF mapping (Fig. 1c and Extended Data Fig. 1n, o), ultraviolet imaging (Extended Data Fig. 1d–m), maturation experiments (Extended Data Fig. 9b, f–q) and TEM of bacteria (Extended Data Fig. 10i)—were not repeated.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

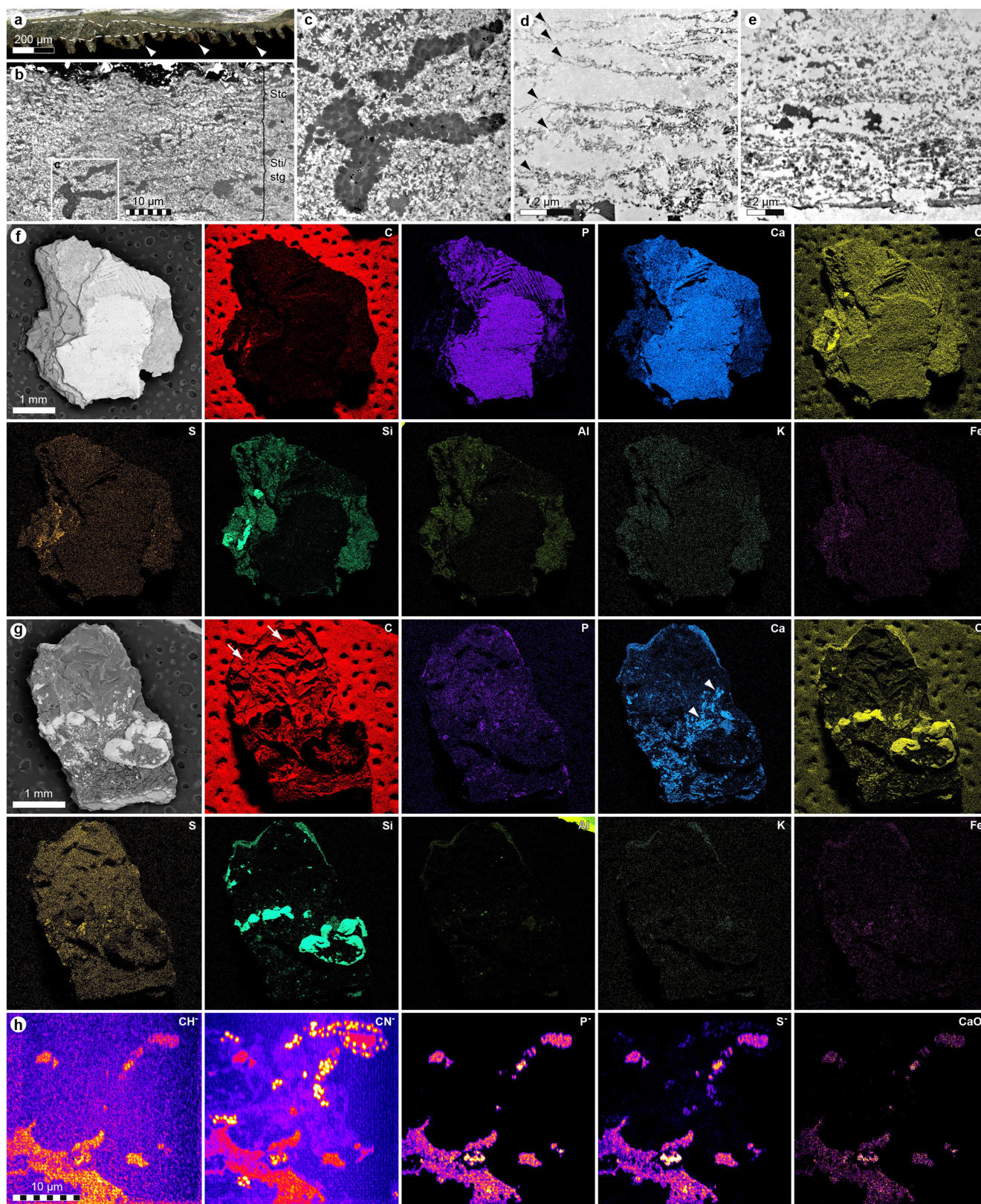
The *Stenopterygius* specimen examined in this study (MH 432) is permanently accessioned in the collections of Urweltmuseum Hauff, Holzmaden, Germany. The data supporting our findings are available from the corresponding author upon reasonable request.

35. Thiel, V. & Sjövall, P. in *Principles and Practice of Analytical Techniques in Geosciences* (ed. Grice, K.) 122–170 (Royal Society of Chemistry, Cambridge, 2015).
36. Ito, S. et al. Acid hydrolysis reveals a low but constant level of pheomelanin in human black to brown hair. *Pigment Cell Melanoma Res.* **31**, 393–403 (2018).



Extended Data Fig. 1 | Regular, ultraviolet light and SRS-XRF images of MH 432. a, Skin compressed onto the diagenetically flattened gastralia basket. **b,** Amorphous adipocere (arrows) external to the gastralia and liver residue (arrowheads) within the abdominal cavity. **c,** Fibrous muscle or connective tissue on the left side of the trunk. **d–m,** Some of the anatomical features, and the stone putty used to reassemble the individual blocks, are enhanced under ultraviolet light. Note the differences in

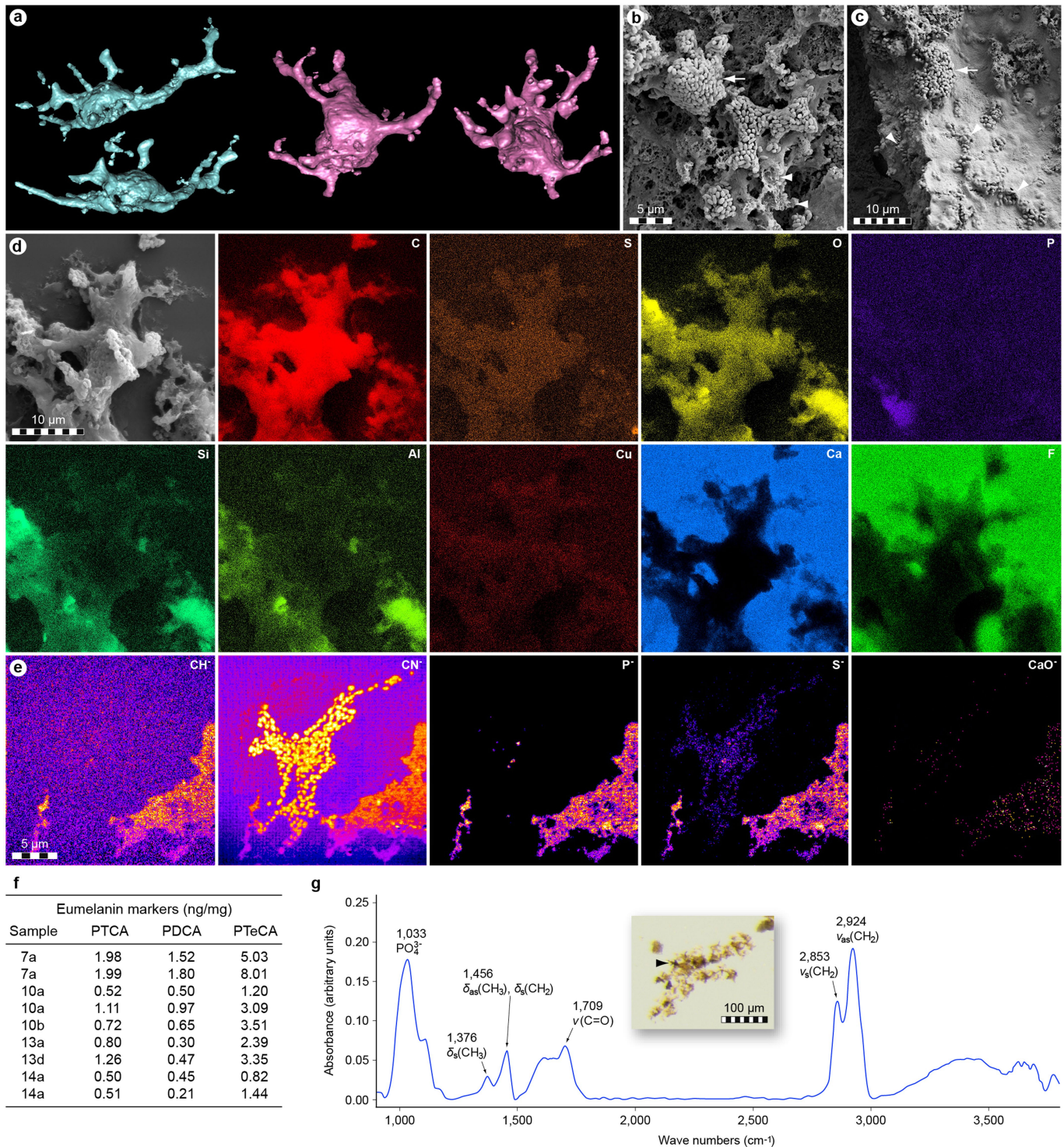
fluorescence between the putty (arrowheads in **f**), internal structures (arrowheads in **g**) and integument (arrowheads in **i**). **n,** SRS-XRF false-colour images showing the spatial distribution of silicon (magenta), phosphorous (green) and copper (blue). **o,** SRS-XRF false-colour images showing the spatial distribution of iron (red), sulfur (yellow) and zinc (blue). The lack of co-localization between copper, zinc and the preserved soft tissues might result from calcium phosphate overprinting.



Extended Data Fig. 2 | See next page for caption.

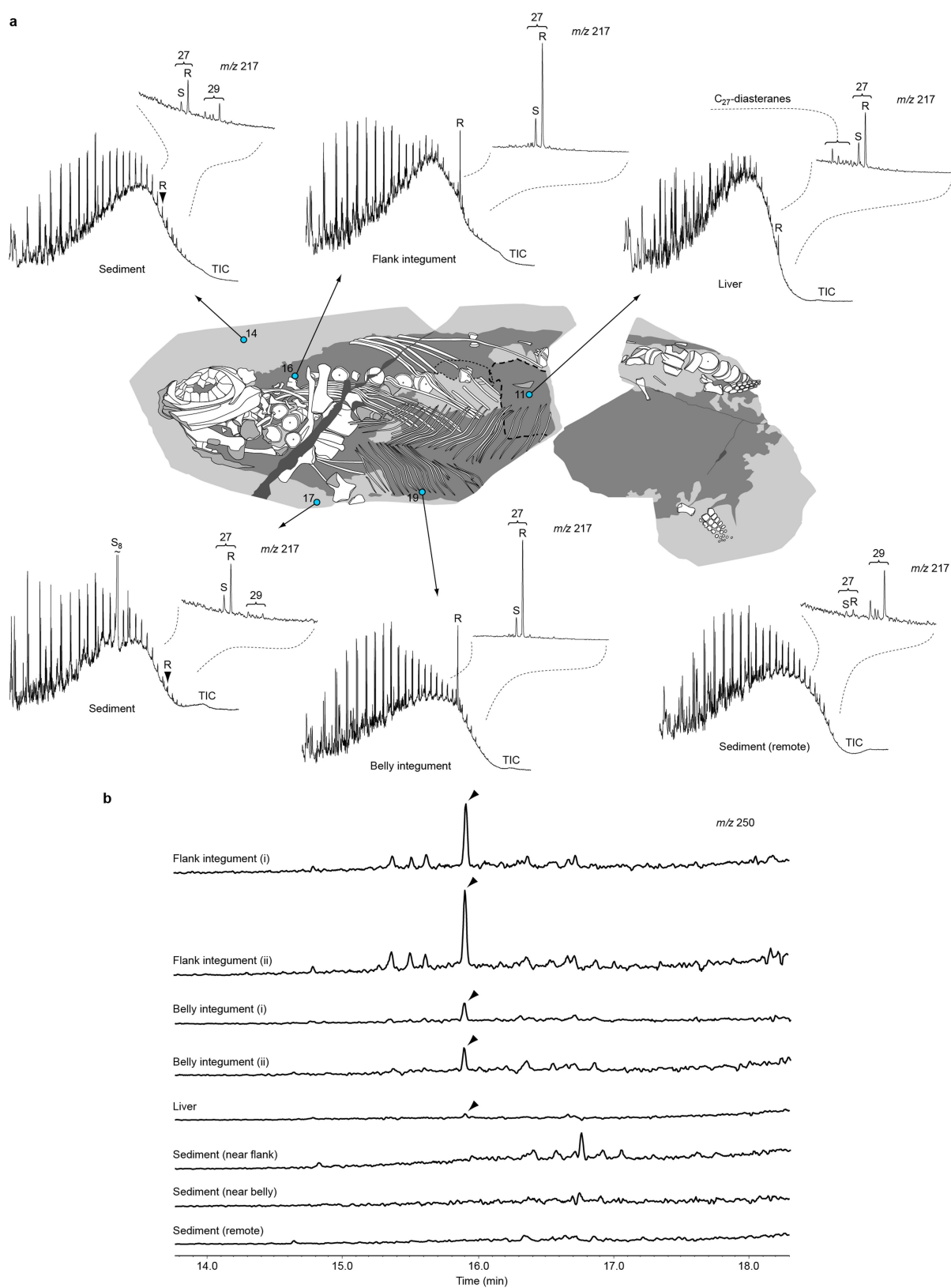
Extended Data Fig. 2 | Light microscopy, TEM, EDX and NanoSIMS data from MH 432 integument. **a**, Transverse section through demineralized epidermis (sample 13a) showing epidermal ridges (arrowheads) and invasive sediments (dashed line). **b**, TEM micrograph through the stratified epidermis (sample 13). Stc, stratum corneum; sti/stg, stratum intermedium and/or stratum germinativum. **c**, Enlargement of a branched epidermal melanophore. **d**, Squamous keratinocytes from between the stratum intermedium and stratum corneum (sample 10); these become progressively flattened towards the exterior surface (top). Arrowheads, cellular envelopes. **e**, TEM section through the fibrous superficial dermis (sample 12a). **f**, Back-scattered electron micrograph and single-element EDX maps of untreated integument in external view. Coloured images illustrate the relative abundance of each element, with higher intensities indicating greater abundance. Note the enrichment

of calcium and phosphorous in the fossilized epidermis and dermis. Intensities from carbon derive primarily from the underlying conductive tape. Al, aluminium (lime); C, carbon (red); Ca, calcium (blue); Fe, iron (violet); K, potassium (turquoise blue); O, oxygen (yellow); P, phosphorous (purple); S, sulfur (orange); Si, silicon (turquoise). **g**, Back-scattered electron micrograph and single-element EDX maps of untreated integument in internal view (colours as in **f**). Note high levels of carbon (arrows) and localized enrichment of calcium (arrowheads) in the blubber and subjacent fibrous tissue. Intensities from silicon and oxygen derive primarily from authigenic silica minerals. **h**, High-resolution NanoSIMS images acquired from demineralized skin showing the distribution of CH^- , CN^- , P^- , S^- and CaO^- (sample 13a). CN^- -rich microbodies are melanosomes (see Extended Data Fig. 3e).



Extended Data Fig. 3 | Three-dimensional visualization and chemistry of MH 432 melanophores. **a**, SRXTM renderings of branched melanophores (sample 18). **b**, FEG-SEM micrograph of a dermal melanophore and adjacent organic matter recovered from demineralized integument (sample 13a). Note the remnant cell body (arrow) and external moulds of disrupted pigment organelles in the polymerized matrix of one dendrite (arrowheads). **c**, FEG-SEM micrograph of a dermal melanophore and adjacent organic matter from demineralized integument (sample 13). Note the clustered melanosomes that represent the cell body (arrow), and dendritic extensions packed with pigment organelles (arrowheads). **d**, Back-scattered electron micrograph and single-element EDX maps of the melanophore in Fig. 3e–g. Coloured images illustrate the relative abundance of each element, with higher intensities indicating greater abundance. Note the enrichment of carbon and, to a lesser extent, sulfur

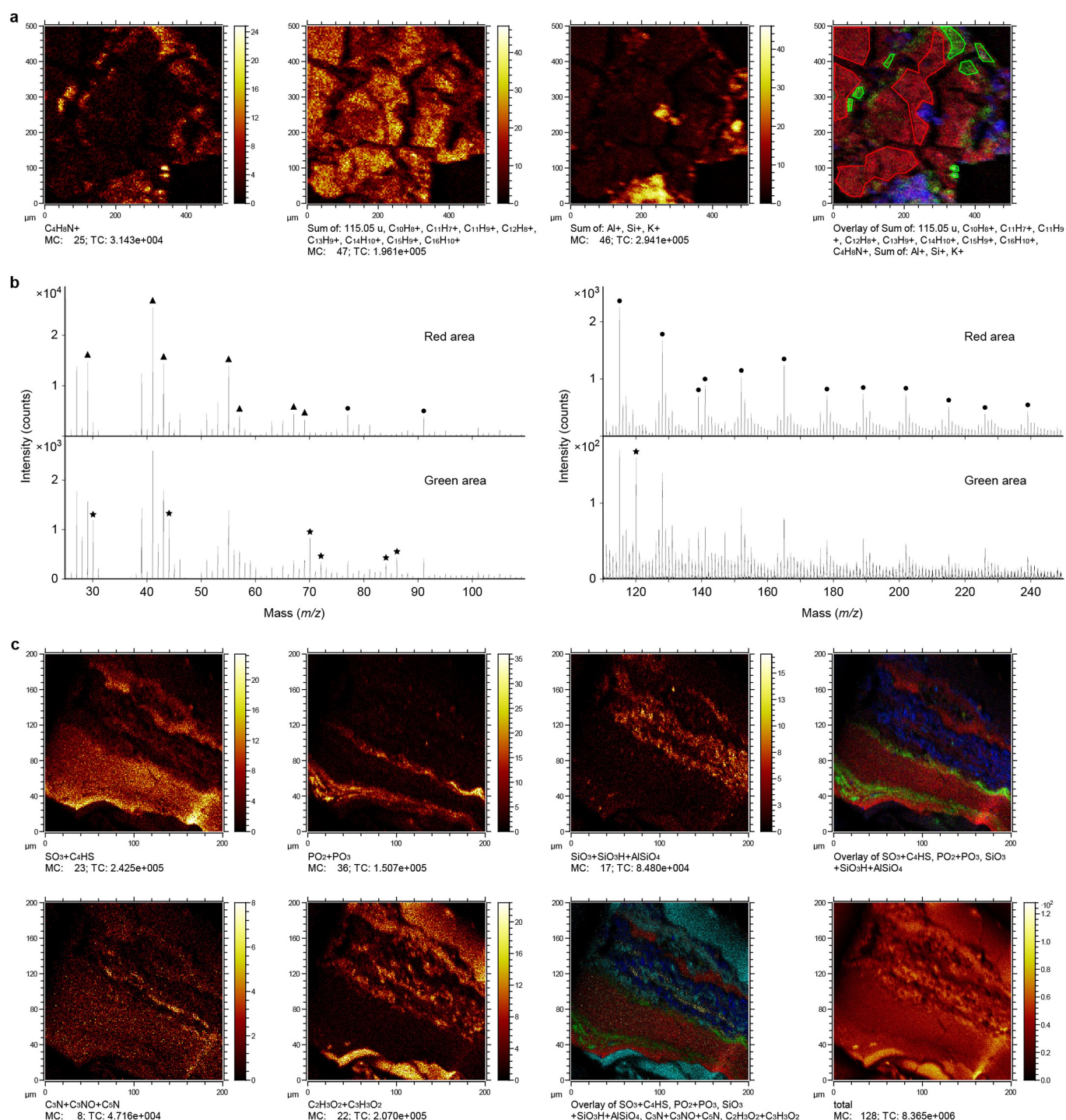
and oxygen in the fossil pigment cell. Intensities from calcium and fluoride derive from the underlying spectrophotometric window (see **g**). Al, aluminium (lime); C, carbon (red); Ca, calcium (blue); Cu, copper (dark red); F, fluoride (green); O, oxygen (yellow); P, phosphorous (purple); S, sulfur (orange); Si, silicon (turquoise). **e**, High-resolution NanoSIMS images of the melanophore in Fig. 3h–k showing the distribution of CH^- , CN^- , P^- , S^- and CaO^- . Note the relatively high levels of CN^- and S^- in the melanosomes, whereas the surrounding matrix also contains measurable amounts of CH^- and P^- . **f**, The alkaline hydrogen peroxide oxidation products PTCA, PDCA and PTeCA (Fig. 3r) from samples 7a (two batches), 10a (two batches), 10b, 13a, 13d and 14a (two batches). **g**, Infrared spectrum from the melanophore in Fig. 3e–g (arrowhead in the inset light microscopy image) showing peaks attributed to hydrocarbons and phosphate.



Extended Data Fig. 4 | See next page for caption.

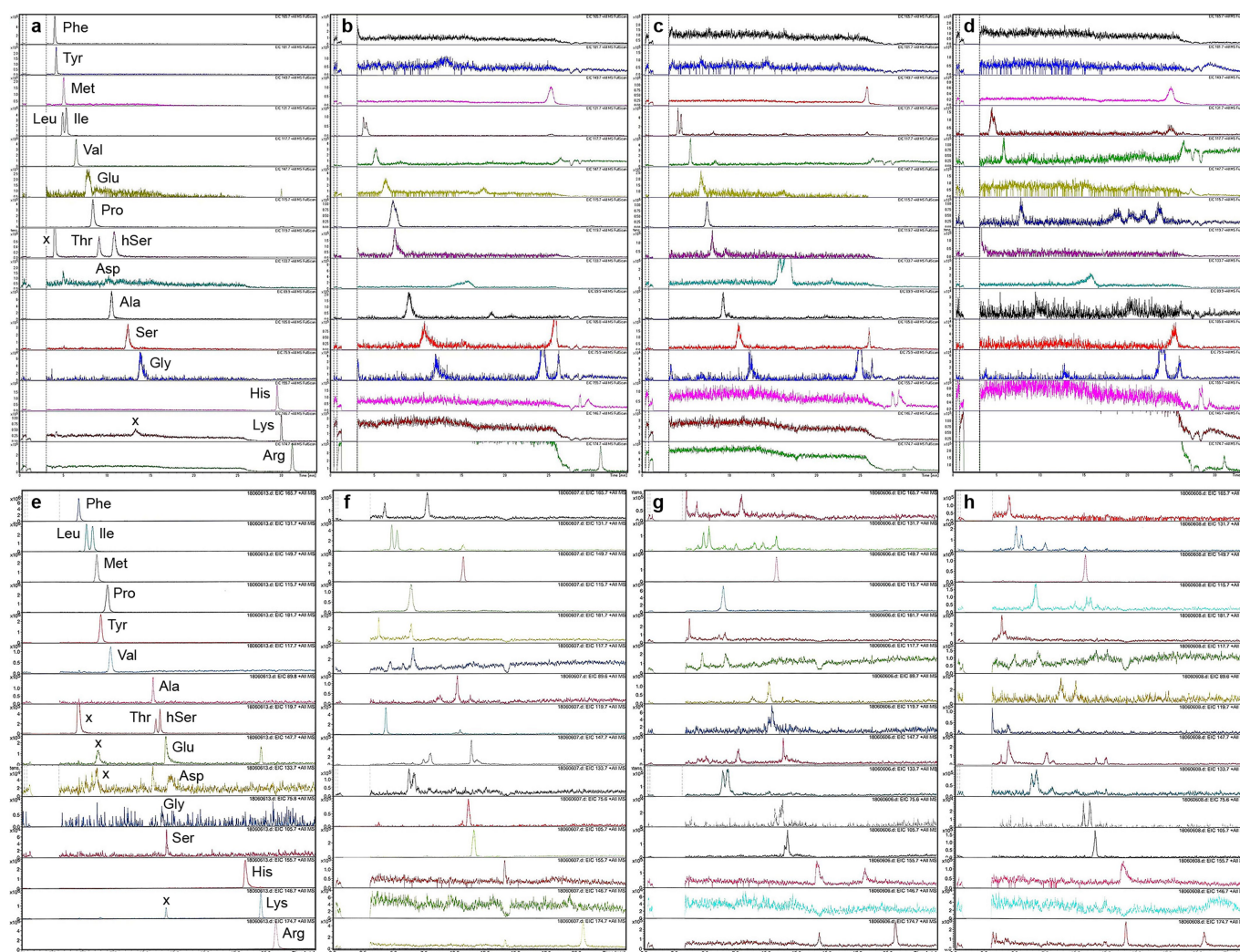
Extended Data Fig. 4 | Pyrolysis data collected at 560 °C for 10 s, from the soft tissues of MH 432. **a**, Py-GC/MS chromatograms obtained from MH 432 (TIC, total ion currents). The prominent peak series in each sample represents homologous *n*-alkenes and *n*-alkanes; chromatograms are normalized to the highest of these peaks. Inset ion chromatograms (*m/z* 217) illustrate the distribution of eukaryote-derived steranes, with 27 and 29 denoting carbon numbers. S and R denote (20S)-5 α ,14 α ,17 α (H) and (20R)-5 α ,14 α ,17 α (H) isomers, respectively. Abundant C₂₇-steranes (cholestanes) in the integument constitute diagenetic products of ichthyosaur cholesterol. The predominant 20R isomer is also indicated in the TICs to illustrate its abundance among the total pyrolysates. C₂₉-steranes (stigmastanes) reflect background sedimentation from algae and/or terrestrial plants. Note the high amount of cholestanes in the

integument and greater abundance of stigmastanes in the host rock. Also note the higher intensities of aromatics (relative to aliphatics), diasteranes (relative to regular steranes) and a stronger unresolved complex mixture in the liver, reflecting original compositional differences and/or enhanced biodegradation. **b**, Py-GC/MS ion chromatograms (*m/z* 250, normalized to sample weight) showing a compound tentatively identified as *n*-octadecadiene (arrowheads, *n*-C_{18:2}). This molecule is interpreted as a pyrolysis product of kerogen-bound *n*-octadecenyl (*n*-C_{18:1}) moieties potentially originating from oleic acid (C_{18:1} ω 9*c*), the most abundant monoenoic fatty acid in extant vertebrates. Note the localized occurrence within the flank integument (where the blubber is best preserved). Replicate sample measurements (denoted by i and ii) are provided to demonstrate reproducibility.



Extended Data Fig. 5 | ToF-SIMS images and spectra from MH 432 integument. a, Images of positive ions (sample 2) that are characteristic of (from left to right) peptides/proteins (C₄H₈N⁺, *m/z* 70), polyaromatic hydrocarbons (C₉H₇⁺, *m/z* 115; C₁₀H₈⁺, *m/z* 128; C₁₁H₇⁺, *m/z* 139; C₁₁H₉⁺, *m/z* 141; C₁₂H₈⁺, *m/z* 152; C₁₃H₉⁺, *m/z* 165; C₁₄H₁₀⁺, *m/z* 178; C₁₅H₉⁺, *m/z* 189; and C₁₆H₁₀⁺, *m/z* 202) and the sedimentary matrix (Al⁺, *m/z* 27; Si⁺, *m/z* 28; and K⁺, *m/z* 39), along with a three-colour overlay image of these ions in which green represents proteinaceous matter, red represents polyaromatic hydrocarbons and blue represents sediment. **b**, Positive ion spectra from selected regions of interest indicated in the three-colour overlay image in **a** (green demarcations highlight areas dominated by proteinaceous matter; red lines frame regions rich in polyaromatic hydrocarbons). Characteristic aliphatic and polyaromatic hydrocarbon peaks are indicated by triangles and circles, respectively,

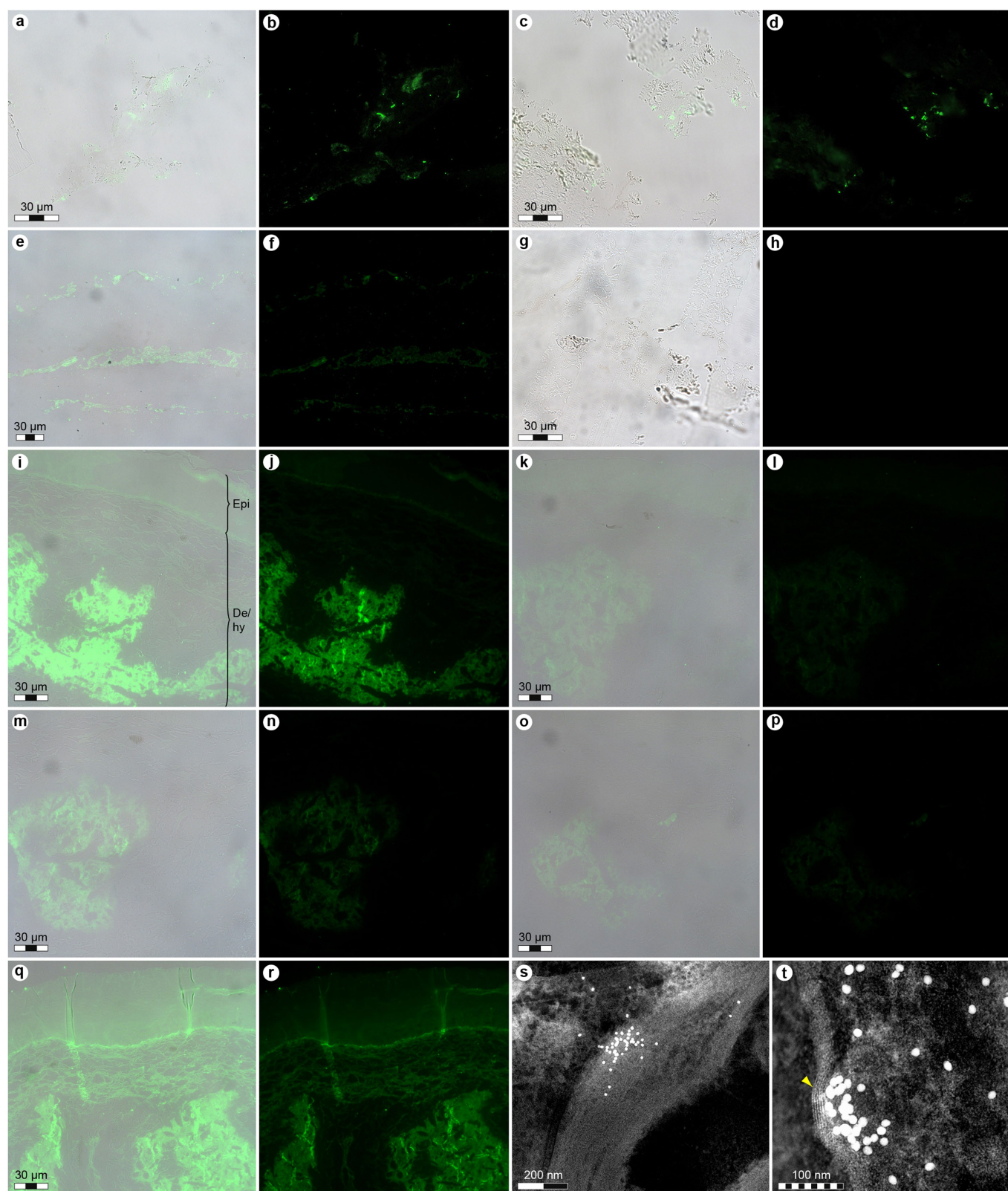
in the top spectrum (red regions of interest), whereas typical protein fragment ions are denoted by stars in the bottom spectrum (green regions of interest; see Supplementary Information). **c**, Negative ion images of sample 13a (see Fig. 5m) representing (top row, from left to right) sulfur-containing materials (SO₃⁻, *m/z* 80 and C₄HS⁻, *m/z* 81), phosphate (PO₂⁻, *m/z* 63 and PO₃⁻, *m/z* 79) and the sedimentary matrix (SiO₃⁻, *m/z* 76; SiHO₃⁻, *m/z* 77; and AlSiO₄⁻, *m/z* 119), together with a three-colour overlay image of these ions in red, green and blue, respectively. The bottom row shows the spatial distribution of ions representing eumelanin (C₃N⁻, *m/z* 50; C₃NO⁻, *m/z* 66; and C₃N⁻, *m/z* 74) and epoxy (C₂H₃O₂⁻, *m/z* 59 and C₃H₃O₂⁻, *m/z* 71), along with a five-colour overlay image featuring sulfur-containing ions (red), phosphate (green), eumelanin (yellow), epoxy (cyan) and sediment (blue); the total ion image is shown on the far right.



Extended Data Fig. 6 | Amino acid analysis data from MH 432.

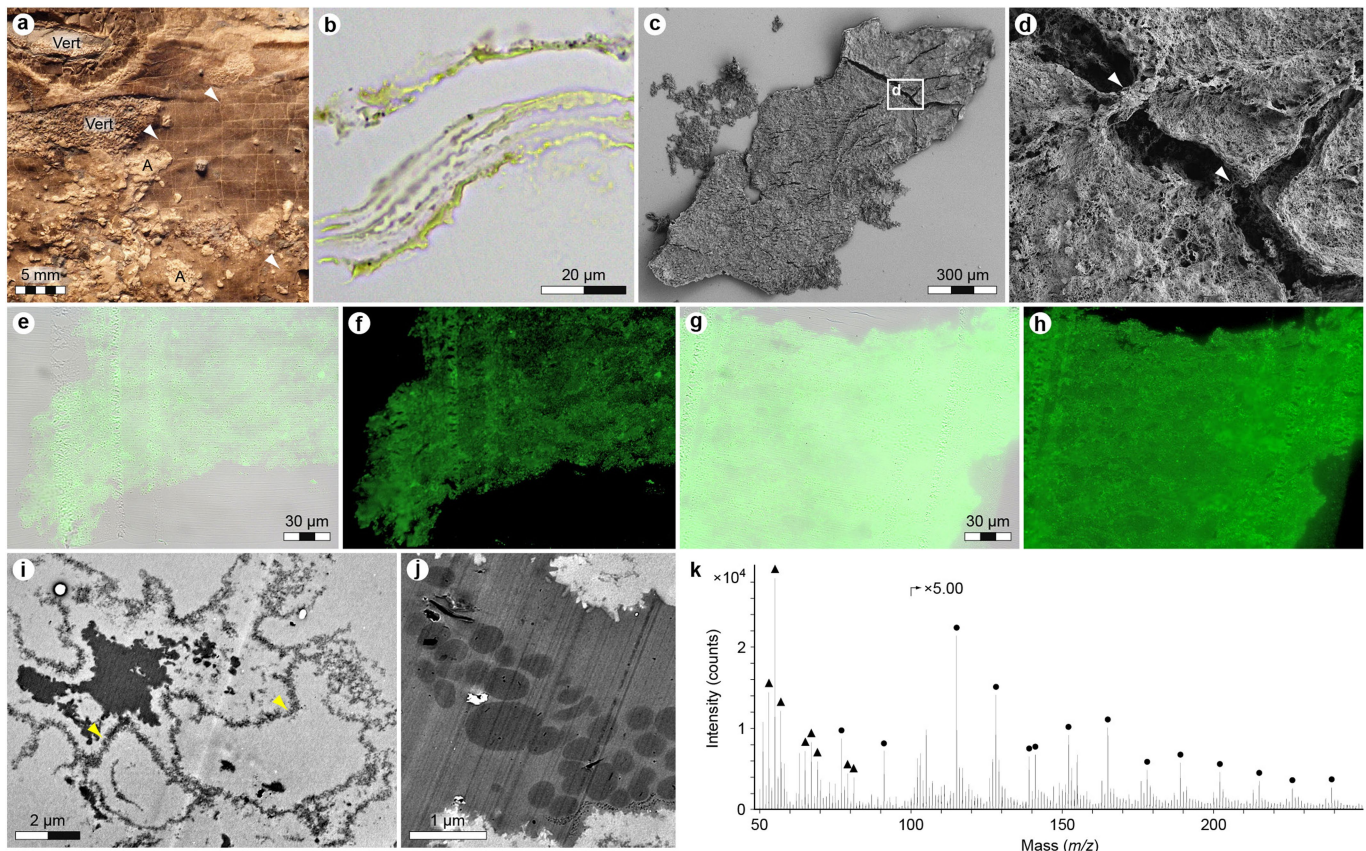
a–d, Liquid chromatography condition I. **a**, Standard amino acids (200 pmol per 4 μ l). **b**, Sample 7a, liver (a 4- μ l portion of 8.51 mg per 170 μ l). **c**, Sample 13d, integument (a 4- μ l portion of 10.27 mg per 206 μ l). **d**, Sample 14a, sediment (a 4- μ l portion of 10.05 mg per 201 μ l). **e–h**, Liquid chromatography condition II. **e**, Standard amino acids (200 pmol per 4 μ l). **f**, Sample 7a, liver (a 5- μ l portion of 8.51 mg per 340 μ l). **g**, Sample 13d, integument (a 5- μ l portion of 10.27 mg per 411 μ l). **h**, Sample 14a, sediment (a 5- μ l portion of 10.05 mg per 402 μ l). Chromatograms for the unstable amino acids asparagine, cysteine,

glutamine and tryptophan were omitted from the figure. Additionally, aspartic acid, methionine and tyrosine were not detected (either owing to their low ionization efficacy or lability). Peaks with retention times similar to those of glycine and serine in the condition-II chromatograms derive from impurities (as demonstrated by the corresponding amino acid peaks in the condition-I chromatograms). Data from condition I were used for alanine, glutamic acid, glycine and serine; data from condition II were used for all other amino acids. The solvent for standard amino acids and extracted samples was 0.1 M HCl. hSer, homoserine. 'x' denotes an impurity.



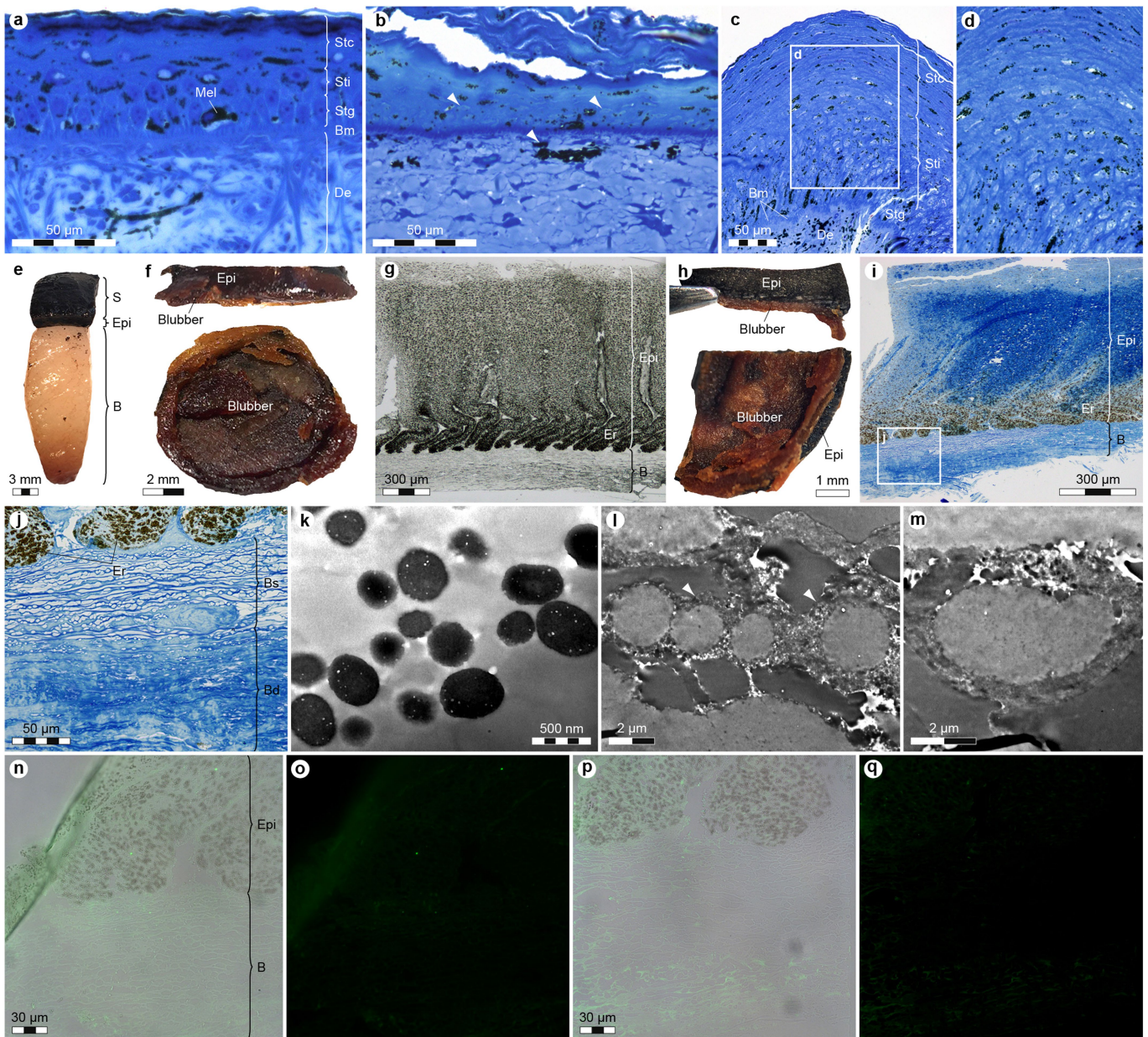
Extended Data Fig. 7 | Immunoreactivity of fossil ichthyosaur and extant leatherback sea turtle skin. a–t, Immunohistochemical staining and immunogold labelling of demineralized MH 432 skin (samples 8 and 12a) (a–h) versus experimentally treated *D. coriacea* skin (i–t), exposed to antibodies raised against *Bos taurus* elastin (a, b), *G. domesticus* actin (c, d), *A. mississippiensis* collagen (e, f, k, l), *G. domesticus* feathers (indicative of β -keratin) (g, h), *G. domesticus* tropomyosin (i, j), *A. mississippiensis* haemoglobin (m, n), *S. camelus* haemoglobin (o, p) and *G. domesticus* α -keratin (q–t). Images in a, c, e, g, i, k, m, o, q show where the antibodies bind to tissue (green) superimposed on transmitted light

images. Fluorescein isothiocyanate fluorescence in b, d, f, h, j, l, n, p, r indicates binding for all antibodies except β -keratin. In *D. coriacea*, the migration from the epidermis to underlying tissues of compounds derived from α -keratin probably reflects the combined effects of decay, compaction and maturation. De/hy, dermis and hypodermis (corresponding to blubber in adult *D. coriacea*); epi, epidermis. s, t, Low-resolution (s) and high-resolution (t) localization of anti- α -keratin antibody tagged with gold to fibrous matter in *D. coriacea* skin. Note the filamentous structures (arrowhead).



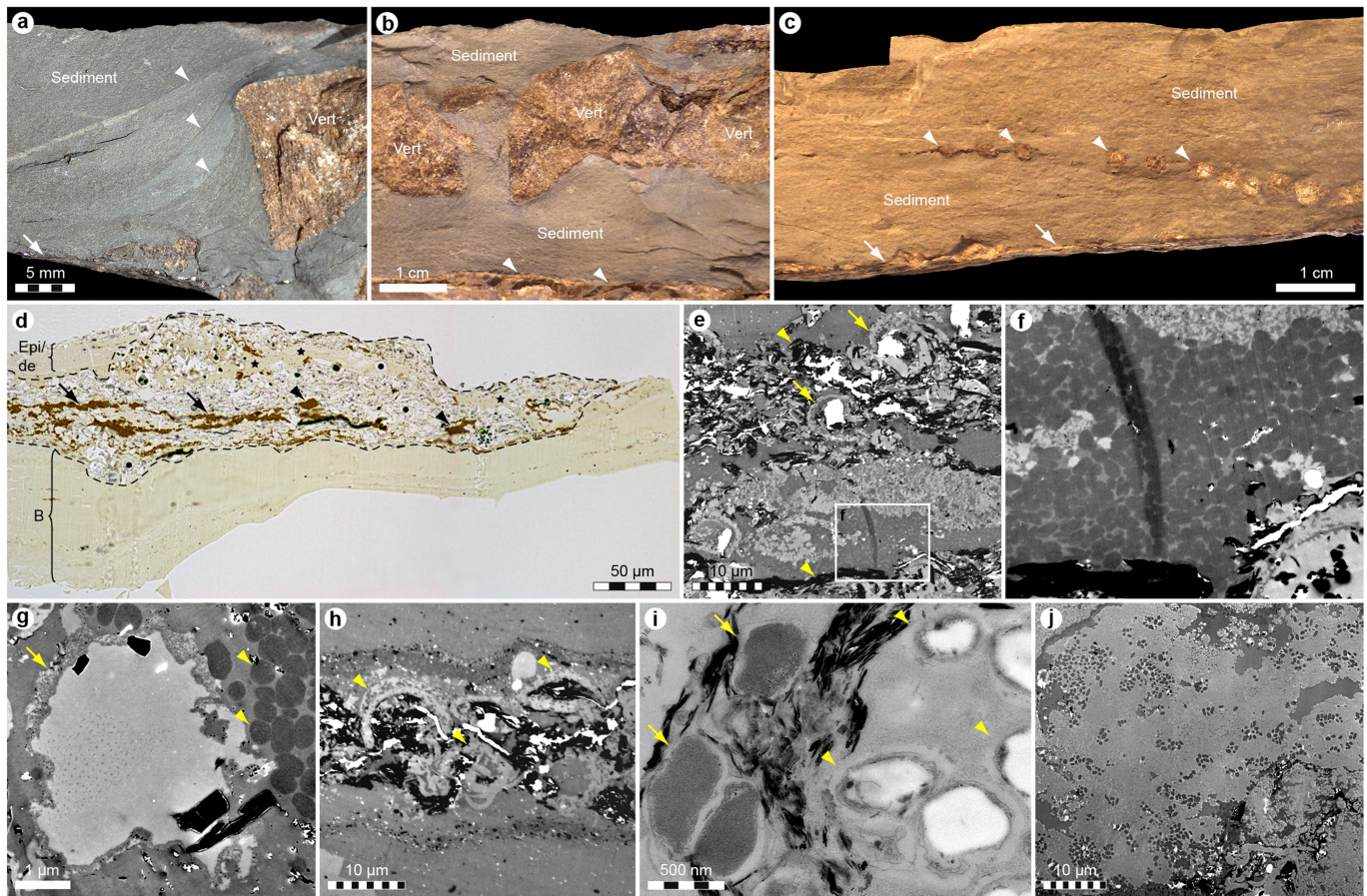
Extended Data Fig. 8 | Ultrastructure and chemistry of the liver of MH 432. **a**, Polygonal surface structure of the red-brown organ trace (arrowheads). Note the irregular patches of phosphatized adipocere (A) overlying the mineralized internal structure, and a vertebral centrum (vert) that has penetrated the decomposing tissue before fossilization. **b**, Light microscopy section through demineralized liver matrix (sample 1), which reveals its layered architecture that is probably produced by diagenetic compaction. **c**, FEG-SEM micrograph of pliable organic matter

released via treatment with EDTA (sample 1). **d**, Enlargement showing a degraded and somewhat fibrous (arrowheads) biomass. **e–h**, Overlay (**e**, **g**) and fluorescent (**f**, **h**) images of demineralized liver material (sample 7) exposed to antibodies raised against *A. mississippiensis* (**e**, **f**) and *S. camelus* haemoglobin (**g**, **h**). **i**, TEM micrograph of probable cellular membranes (arrowheads). **j**, TEM micrograph showing a dense melanosome cluster (sample 1). **k**, Positive ion ToF-SIMS spectrum with peaks characteristic of aliphatics (triangles) and polyaromatics (circles; sample 1).



Extended Data Fig. 9 | Experimental maturation of extant skin and subcutis to simulate the effects of diagenesis. **a**, Stained light microscopy section through the scaly carapace integument from a juvenile *D. coriacea*. Bm, basement membrane; de, dermis; mel, melanophore; stc, stratum corneum; stg, stratum germinativum; and sti, stratum intermedium. **b**, Autoclave-treated carapace integument from a juvenile *D. coriacea* showing flattened keratinocytes and melanophores (arrowheads). **c**, Scaleless carapace integument of an adult *D. coriacea* revealing multiple layers of stratified squamous keratinocytes. Note the greater thickness of the stratum intermedium in the adult relative to the juvenile individual (compare with **a**). **d**, Enlargement of the stratum corneum and stratum intermedium. **e**, Melanized *P. phocoena* body integument. B, blubber (dermis and subcutis); epi, epidermis; s, outer skin surface (in oblique aspect). **f**, Side (top) and internal (bottom) views of artificially compressed

P. phocoena integument. **g**, Light microscopy section through artificially compressed *P. phocoena* integument showing the condensed blubber layer (compare with **e**). Er, epidermal ridge. **h**, Side (top) and internal (bottom) views of *P. phocoena* integument following autoclave experiments. **i**, Stained light microscopy section through autoclave-treated integument of *P. phocoena*. **j**, Enlargement of the loosely packed superficial blubber (Bs)—a possible entry for microbes (compare Extended Data Fig. 10d)—and dense deeper blubber (Bd). **k**, Melanosomes in experimentally treated epidermis of *P. phocoena*. **l**, **m**, Shrunken, membrane-bound (arrowheads) adipocytes (or lipid vesicles) in experimentally treated blubber of *P. phocoena* (compare Extended Data Fig. 10g). **n–q**, Overlay (**n**, **p**) and fluorescent (**o**, **q**) images of experimentally treated integument of *P. phocoena*, exposed to antibodies raised against *G. domesticus* α -keratin (**n**, **o**) and *G. domesticus* tropomyosin (**p**, **q**).



Extended Data Fig. 10 | Taphonomy of MH 432. **a**, Cross-section through the main rock slab (left side of posterior termination) in original geological orientation, showing sediment infill between the integument (arrow) and a sectioned vertebra (vert). Saturated mud (arrowheads) encased the carcass following gravitational collapse of the backbone. **b**, Natural break through the main slab (centre of posterior termination) exposing intrusive sediment infill into the body cavity before disarticulation of the vertebral column. Arrowheads indicate the liver residue. **c**, Cross-section through the main slab (right side of posterior termination). Note the invasive sediment covering the residual soft parts (arrows) and dorsal ribs from the right side of the body (arrowheads). **d**, Light microscopy section through demineralized integument (sample 13a) with clay minerals and inferred bacteria (delimited by dashed line) penetrating between the phosphatized epidermis and/or dermis (Epi/de)

and polymerized blubber (B). Skin (stars) and melanophores (arrowheads) occur along with hollow structures interpreted as bacterial cellular bodies (circles) and massed melanosomes (arrows). **e**, TEM micrograph of sample 13a showing clay minerals (arrowheads) and bacterial cells (arrows). **f**, Enlarged melanosome concentration produced by microbially mediated skin reduction. **g**, Adipocyte or lipid vesicle (compare Extended Data Fig. 9l, m), or microbe (compare with **i**), with well-developed cellular membrane (arrow) and adjacent melanosomes (arrowheads). **h**, Collapsed thick-walled bacterial cells (arrowheads) that suggest microorganismal infestation before fossilization and diagenetic compaction (sample 13a). **i**, Comparative image of extant bone-boring bacteria (arrows) showing retention of cellular membranes (arrowheads) after removal of internal contents. **j**, Decomposed skin of MH 432 (sample 13).

Stochastic synaptic plasticity underlying compulsion in a model of addiction

Vincent Pascoli¹, Agnès Hiver¹, Ruud Van Zessen¹, Michaël Loureiro¹, Ridouane Achargui¹, Masaya Harada¹, Jérôme Flakowski¹ & Christian Lüscher^{1,2*}

Activation of the mesolimbic dopamine system reinforces goal-directed behaviours. With repetitive stimulation—for example, by chronic drug abuse—the reinforcement may become compulsive and intake continues even in the face of major negative consequences. Here we gave mice the opportunity to optogenetically self-stimulate dopaminergic neurons and observed that only a fraction of mice persevered if they had to endure an electric shock. Compulsive lever pressing was associated with an activity peak in the projection terminals from the orbitofrontal cortex (OFC) to the dorsal striatum. Although brief inhibition of OFC neurons temporarily relieved compulsive reinforcement, we found that transmission from the OFC to the striatum was permanently potentiated in persevering mice. To establish causality, we potentiated these synapses *in vivo* in mice that stopped optogenetic self-stimulation of dopamine neurons because of punishment; this led to compulsive lever pressing, whereas depotentiation in persevering mice had the converse effect. In summary, synaptic potentiation of transmission from the OFC to the dorsal striatum drives compulsive reinforcement, a defining symptom of addiction.

All addictive drugs target the mesolimbic dopamine system¹. Through distinct cellular mechanisms, these drugs increase dopamine levels² even in the absence of a reward-prediction error, resulting in an excessive learning signal³. This may lead to loss of control, such that some individuals will shift to compulsive drug intake^{4,5}, used by some to define addiction^{6–8}. About 20% of users of addictive substances such as cocaine, heroin and amphetamines eventually fulfil this diagnostic criterion⁹.

The neural correlate for compulsive reinforcement is poorly understood but an imbalance in the systems that control goal-directed and habitual actions has been implicated^{5,10,11}, possibly driven by an activity shift from the ventro-medial to dorso-lateral striatum¹². The formation of stimulus–response associations may trigger motor programs in the dorso-lateral striatum that favour habitual drug use¹². Alternatively, addicts may suffer from a failure of ‘top-down’ inhibition of stimulus–response associations, a function attributed to the medial prefrontal cortex^{13,14}. Finally, drugs may perturb goal-directed outcome-representation ascribed to the OFC. Indeed, pharmacological inhibition of the dorsal striatum reduced cocaine-seeking behaviour under punishment paradigms¹⁵, and optogenetic stimulation of the prelimbic cortex has the same consequence¹⁴. Compulsive self-administration of cocaine is associated with enhanced activity in the OFC and inhibition of the OFC reduces compulsive reinforcement^{16,17}. Moreover, the function of the OFC is disrupted after withdrawal from cocaine self-administration in rats^{18–20}.

Although the OFC and the striatum emerge as hubs for compulsive reinforcement, the cellular substrate that maintains reward-seeking behaviours despite negative consequences remains unknown.

Because an increase in the levels of mesolimbic dopamine is the defining commonality of addictive drugs, we implemented a model in which the mouse presses a lever to optogenetically activate the ventral tegmental area (VTA) dopaminergic neurons (optogenetic dopamine-neuron self-stimulation; oDASS), which mimics drug-induced circuit-wide adaptations^{16,21}. Here we identify a cellular correlate of

compulsive reinforcement, by introducing a punishment once oDASS was established.

Compulsive self-stimulation of dopamine neurons

We expressed channelrhodopsin-2 (ChR2) bilaterally in dopaminergic neurons of the VTA and implanted an optic fibre that was aimed at the midbrain (oDASS mice, see Methods and Extended Data Fig. 1). When a mouse pressed the active lever, laser stimulation (30 bursts of 5 pulses of 4 ms at 20 Hz) began after a 5-s delay. Every four days, the number of lever presses required to trigger stimulation was increased to reach a final fixed ratio of three (FR3). All mice quickly reached a maximum of 80 laser stimulations in less than one hour (Fig. 1a, b).

Following acquisition, the perseverance of oDASS despite electric foot shocks was used to evaluate compulsivity. The shock was delivered every third completed fixed ratio schedule and its intensity (500 ms, 0.25 mA) was sufficient to suppress lever pressing for sucrose reward¹⁶. This punishment reduced the laser-stimulation rate (Fig. 1b), albeit with high variability between individual mice. Some mice almost stopped responding, whereas others kept obtaining the maximum oDASS, taking only slightly more time (Extended Data Fig. 2). For the 109 mice, the histogram for the oDASS rate was unimodal during the baseline sessions but became bimodal by the end of the fourth punished session (Fig. 1c). A clustering method on the entire behavioural dataset revealed two distinct classes (Extended Data Fig. 3): mice with a small decrease in oDASS rate during punished sessions ($n = 66$) and mice that strongly decreased oDASS ($n = 43$), which we called perseverers and renouncers, respectively. Because we found similar proportions of male and female mice in both clusters, all subsequent experiments were carried out on both sexes (Fig. 1d). When subjected to additional punishment sessions, mice remained in the original cluster (Extended Data Fig. 4a). Moreover, there was no correlation between perseverance and baseline oDASS rate or motivation for oDASS (Fig. 1d, e and Extended Data Fig. 4b, c). Thus, in about 60% of the mice, the burst activity elicited by oDASS

¹Department of Basic Neurosciences, Faculty of Medicine, University of Geneva, Geneva, Switzerland. ²Clinic of Neurology, Department of Clinical Neurosciences, Geneva University Hospital, Geneva, Switzerland. *e-mail: Christian.Luscher@unige.ch

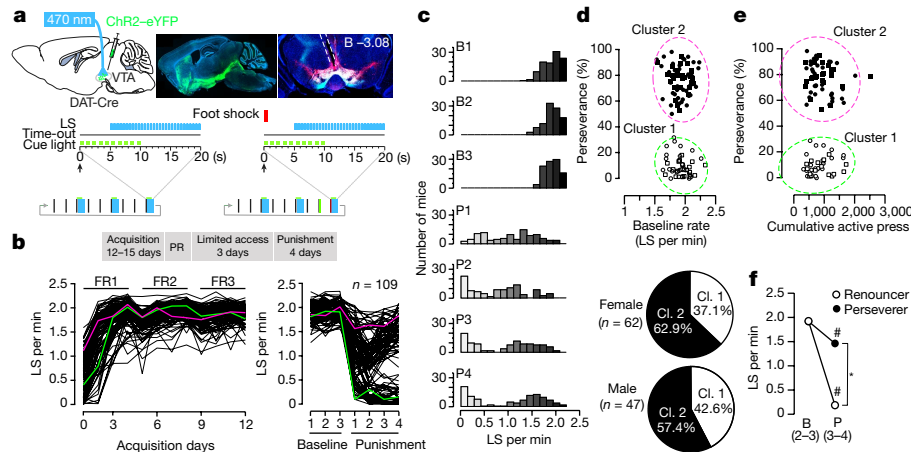


Fig. 1 | Perseverance of oDASS despite punishment. **a**, Schematic of optic fibre placement above the VTA of DAT-Cre mouse infected with AAV5-EF1 α -DIO-ChR2-eYFP (left), image of parasagittal oblique slice and tyrosine hydroxylase staining of a midbrain coronal slice (right). **B**, bregma; values indicate coordinates from bregma. Mice pressed a lever during a sequence consisting of three FR3 trials (a total of 9 lever presses) to self-stimulate. Completion of a FR3 trial triggered a cue light and 5 s later the laser stimulation (LS, 30 bursts composed of 5 short laser pulses, 4-ms width at 20 Hz) as well as a time-out period during which lever presses had no consequences. During punished sessions for every third trial, the cage light was turned on for 1 s at the second press (number 8) of the FR3 and the final press (number 9) triggered a foot shock (0.25 mA, 500 ms). **b**, Mice were subjected to punishment sessions after 12 days of acquisition (maximum of 80 oDASS per day), a progressive ratio session and 3 days of limited access (maximum of 40 oDASS per day). PR, progressive ratio. oDASS rate (LS per min) during acquisition (left), baseline and punished sessions (right) for all mice ($n = 109$ mice). Green and pink lines identify the examples of two mice displayed in

Extended Data Fig. 2. **c**, Histograms of the proportion of mice binned by oDASS rate during baseline (B1–B3) and punished sessions (P1–P4) ($n = 109$ mice). **d**, Perseverance (rate of oDASS during punished sessions 3–4, normalized to baseline) as a function of the baseline rate for male and female mice, showing the two clusters identified in Extended Data Fig. 3. Pie charts show the proportion of male and female mice in clusters 1 and 2 ($n = 109$ mice). **e**, Perseverance as a function of oDASS motivation, measured by the cumulative number of active presses during a progressive ratio schedule session ($n = 98$ mice). **f**, oDASS rate (mean \pm s.e.m.) during the two last sessions of baseline and punished for renouncing and persevering mice (analysis of variance (ANOVA) followed by two-sided t -test: $t_{107} = 0.11$, $P > 0.99$ and $t_{107} = 33.69$, $P < 0.0001$; * $P < 0.05$ for persevering versus renouncing mice for the punished session; $n = 66$ and 43 mice, respectively; $t_{42} = 50.24$, $P < 0.0001$ and $t_{65} = 16.26$, $P < 0.0001$; # $P < 0.05$; baseline compared to punished sessions for renouncing and persevering mice). See Supplementary Table 1 for complete statistics. **a**, Line drawing modified from Paxinos and Franklin⁴⁴, copyright © 2007.

was sufficient to stochastically induce perseverance despite negative consequences.

At baseline, the average delay between completion of FR sequences and the initiation of a subsequent trial was less than 10 s (Extended Data Fig. 4d). Once punishment was introduced, renouncers showed a strong increase in the delay to initiate the next sequence. After a completed FR trial without foot shock the delays became shorter, suggesting that continuous updating of reward value in light of the preceding outcome occurred.

The longer delays led to a reduced oDASS rate over the entire session that was different between the two clusters, dropping to less than 20% of baseline in renouncers, compared to about 80% in persevering mice (Fig. 1f).

The OFC projects to the striatum

Following a previously published screen¹⁶, we next mapped the circuit that originates in the lateral OFC. An adeno-associated virus (AAV8-hSyn-chrimson-tdTomato) that was injected in the OFC anterogradely labelled fibre terminals in the centro-ventral part of the dorsal striatum, all along the rostro-caudal axis (Fig. 2a and Supplementary Video 1). Conversely, retrograde labelling by seeding CTB-555 in the striatum stained neurons in layers II, III and V of the OFC (Extended Data Fig. 5a).

In acute brain slices, terminal stimulation evoked large currents (400–1,000 pA) in all neurons that are located in the centro-ventral part of the striatum; these currents had both α -amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor (AMPA) and N -methyl-D-aspartic acid receptor (NMDAR) components (Fig. 2b). In other striatal regions, connection rates were low and current amplitudes small. When sequentially recording excitatory and inhibitory transmission, we found that the onset of inhibitory currents lagged behind excitatory currents by 2–5 ms, consistent with the existence of a feed-forward circuit²² (Extended Data Fig. 5b). When resolving the cell-type specificity using

dopamine receptor D1R-Cre, D2R-Cre and parvalbumin-Cre mouse lines, we observed strong, converging afferents to both subtypes of spiny projection neurons (SPNs) and sparse connections onto parvalbumin interneurons (Fig. 2c, Extended Data Fig. 5c and Supplementary Video 2). Together, these experiments highlight a very strong excitatory projection from the OFC to the striatum, onto both D1R- and D2R-SPNs.

Terminal activity of OFC–striatum

We infected mice with the fluorescent calcium sensor AAV-DJ-CamKII-GCaMP6m in the OFC and placed a photometry fibre in the striatum (Fig. 3a). In baseline sessions, calcium signals decreased around lever presses in all animals (Fig. 3b and Extended Data Fig. 6a). During punished sessions, a similar decrease was observed in renouncing mice, whereas in persevering mice the calcium signal started to increase just before the lever was pressed (Fig. 3b and Extended Data Fig. 4b). In persevering mice, activity during baseline and punished sessions was therefore markedly different (Fig. 3c, d). Unpredictable foot shocks increased activity in both renouncing and persevering mice, with activity peaking after the onset of the foot shock (Extended Data Fig. 6c). Overall, the inversion of the calcium signal—which correlates with perseverance—implicates activity of the OFC–striatum pathway in compulsion.

Time-locked inhibition of OFC

We next transiently inhibited OFC neurons during oDASS to curb the increased activity observed in persevering mice. Amber light (593 nm) activated eArchT3.0 (Fig. 3e) and suppressed action potentials in slices (Fig. 3f). Inhibition in vivo immediately after the lever press that triggered the punishment-predictive cue induced a pause that eventually reduced the oDASS rate (Fig. 3g). We next inhibited OFC neurons after the completion of a sequence, which yielded a temporal profile that was similar to the profile of renouncing mice and an oDASS rate

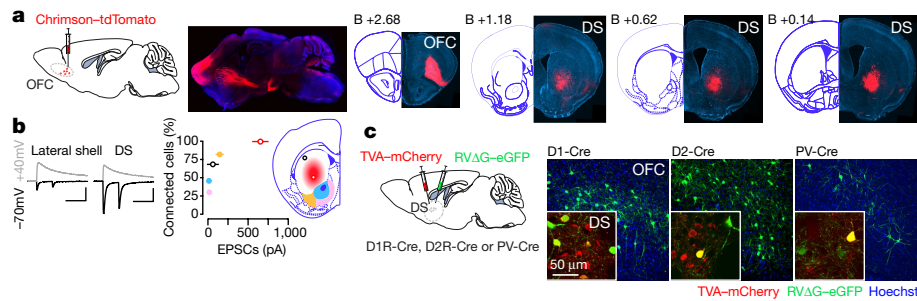


Fig. 2 | The OFC projects to the ventro-central part of the dorsal striatum. **a**, Schematic of the preparation for anterograde tracing from the OFC in wild-type mice. Representative sagittal and coronal images of injection site and terminals at four rostro-caudal coordinates (from bregma) of the striatum (repeated in 10 mice). DS, dorsal striatum. **b**, Left, example traces for optogenetically evoked OFC–striatum EPSCs recorded at +40 mV and -70 mV. Scale bars, 20 ms, 100 pA. Right, functional connectivity between OFC and striatal subregions: nucleus accumbens core (blue, $n = 13$ cells), medial and lateral shell (pink and yellow, $n = 23$ and 34 cells, respectively), dorso-lateral striatum and ventro-central

striatum (black and red, $n = 16$ and 112 cells, respectively). Data are mean \pm s.e.m. **c**, Schematic of retrograde tracing from specific cell types of the striatum using a rabies virus injected in transgenic mouse lines expressing Cre under the control of the dopamine D1, D2 receptor or parvalbumin (PV). Confocal pictures show retrogradely infected neurons in the OFC and starter cells in the striatum at high magnification (insets). Experiments were repeated in $n = 3, 3$ and 4 mice for D1R-Cre, D2R-Cre and PV-Cre mice, respectively. **a, c**, Line drawings modified from Paxinos and Franklin⁴⁴, copyright © 2007.

below 45%. The same intervention during a baseline session had no consequences for the oDASS rate (Fig. 3h). Similarly, inhibition of OFC neurons before every FR trial led to a reduction of perseverance (Extended Data Fig. 6d). In renouncing mice, inhibition of OFC neurons had little consequences for the already long delays between lever presses (Extended Data Fig. 6e).

These results suggest that OFC activity is required to invigorate compulsive oDASS, especially at the time when mice engage in the next sequence. However, the effect was transient and perseverance

returned the following day (Fig. 3i), which is why we next searched for a long-lasting alteration in the OFC output.

Plasticity at OFC–striatum synapses

We performed ex vivo recordings in brain slices from oDASS mice that expressed *Chrimson* in the OFC with on-the-fly identification of SPNs²³ (Fig. 4a), which was confirmed post hoc (Fig. 4b). The ratio of AMPAR to NMDAR excitatory postsynaptic current amplitudes was significantly higher in both D1R- and D2R-SPNs in persevering

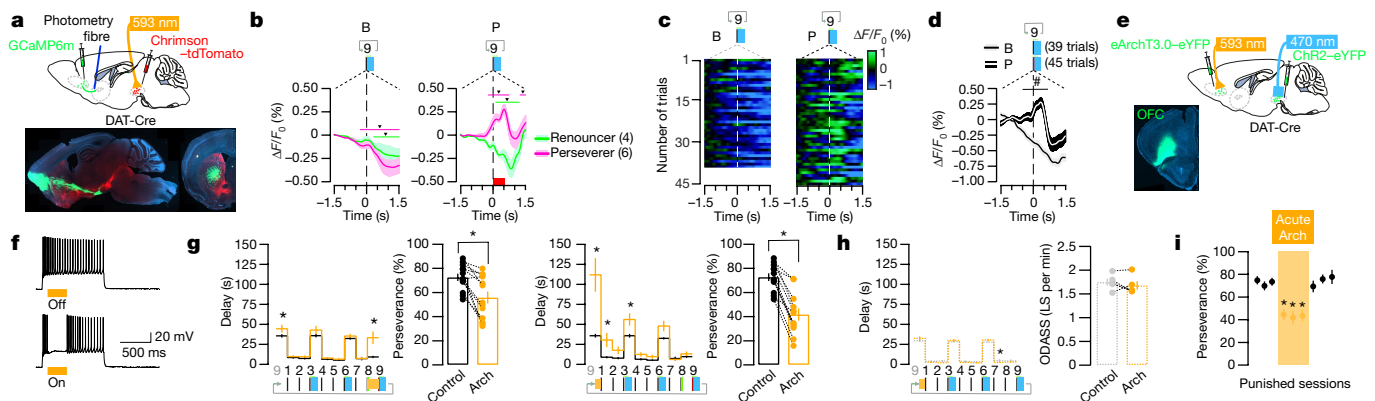


Fig. 3 | Observation and manipulation of OFC–striatum activity during oDASS. **a**, Schematic of the preparation of activity recordings of OFC terminals in the striatum with photometry (top) and image of a parasagittal oblique section showing GCaMP6m expression in the OFC and Chrimson–tdTomato in the VTA (repeated in 10 mice, bottom). **b**, Calcium signal ($\Delta F/F_0$) around the active lever press number 9 completing a sequence during baseline and punished sessions for renouncing and persevering mice. Red block indicates duration of electric shock. Downward triangles and green and pink lines indicate time window with significant deviation from baseline, shaded area represents s.e.m.; $n = 4$ and 6 mice. **c**, Trial activity map of calcium signals ($\Delta F/F_0$) around the active press completing the third FR3 during baseline (left) and punished (right) sessions for a persevering mouse. **d**, Group data for **c**. *Black line indicates the time window with a significant difference between punished and baseline trials using a two-sided permutation test; $n = 39$ and 45, respectively. **e**, Schematics and image of a mouse brain infected with eArchT3.0–eYFP (Arch) in the OFC and with ChR2–eYFP in the VTA (repeated in 20 mice). **f**, Activation of eArchT3.0 inhibits action potentials induced by a current step in an OFC slice. **g**, In persevering mice, OFC inhibition at specific time points during punished oDASS sessions specifically modifies the delay to engage the next action (ANOVA,

followed by two-sided t -test: $*P < 0.05$ comparing delays during punished sessions for control and eArchT3.0 stimulation). Perseverance changed as a consequence of eArchT3.0 stimulation (two-sided paired t -test: $t_{12} = 4.51$, $*P = 0.0007$, $n = 13$ mice for control and eArchT3.0 stimulation at punishment-predicted cue; $t_{11} = 8.91$, $*P < 0.0001$, $n = 12$ mice for control and eArchT3.0 stimulation after punished oDASS). **h**, Delay to engage the next action during baseline with or without OFC inhibition, with eArchT3.0 stimulation after oDASS (ANOVA followed by two-sided t -test: $*P < 0.05$ comparing delays during baseline sessions for control and eArchT3.0 stimulation). Inhibition using eArchT3.0 during baseline sessions had no consequences (two-sided paired t -test: $t_4 = 0.50$, $P = 0.64$, for oDASS rate, $n = 5$). **i**, During additional punished sessions without renewal of the intervention (inhibition after punished oDASS), the effect on perseverance was not maintained (ANOVA followed by Dunnett's test, for three consecutive comparisons: $q_{11} = 7.31$, $*P = 0.0001$; $q_{11} = 6.79$, $*P = 0.0002$; $q_{11} = 6.15$, $*P = 0.0004$; and $q_{11} = 0.96$, $P = 0.84$; $q_{11} = 0.67$, $P = 0.96$; $q_{11} = 1.23$, $P = 0.67$ for every punished versus eArchT3.0 session and punished versus recovery session, respectively; $n = 12$ mice). Data are mean \pm s.e.m. for all panels. See Supplementary Table 1 for complete statistics. **a, e**, Line drawings modified from Paxinos and Franklin⁴⁴, copyright © 2007.

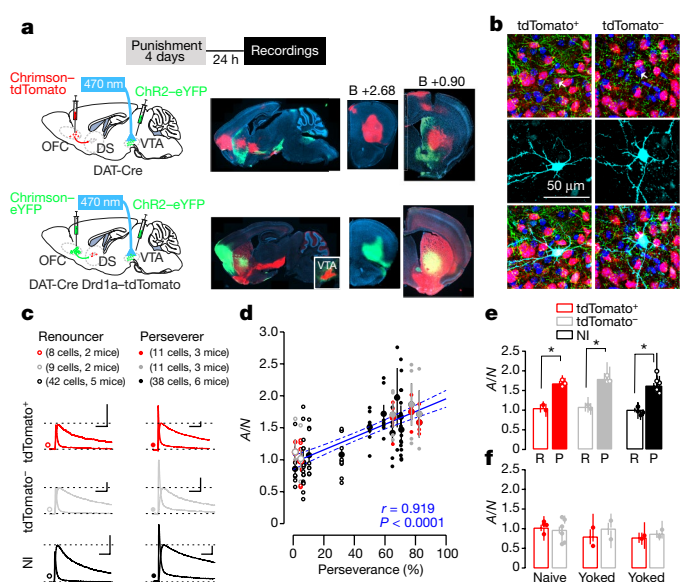


Fig. 4 | Correlation between plasticity of OFC–striatum synapses and compulsive oDASS. **a**, Schematic of the preparation for ex vivo recordings of OFC–striatum synapses. Recordings were performed 24 h after the fourth punishment session in slices from DAT-Cre (encoded by *Dat^{Cre}*) mice or from *Dat^{Cre}Drd1a^{tdTomato}* mice (VTA infected with Chr2-eYFP) expressing Chrimson–tdTomato or eYFP in the OFC, respectively. Parasagittal oblique and coronal sections show infection in the OFC and terminals in the striatum (repeated in six mice). Note that VTA infection in *Dat^{Cre}Drd1a^{tdTomato}* mice is hidden by tdTomato⁺ neurons from the striatum projecting to the midbrain. Inset shows VTA from the same animal, but from a more-medial section (repeated in three mice). **b**, In *Dat^{Cre}Drd1a^{tdTomato}* mice, recorded neurons were filled with biocytin for the identification of SPN subtypes (repeated in 39 slices from 6 mice). **c**, Average of 10 sweeps for AMPAR EPSCs in the presence of D-2-amino-5-phosphonovaleric acid (D-AP5) (50 μ M) and NMDAR EPSCs isolated by subtraction for renouncing and persevering mice, with or without SPN identification with tdTomato. NI, not identified. Scale bars, 200 pA, 50 ms. **d**, AMPAR/NMDAR (A/N) ratio for every neuron as a function of perseverance, per animal and correlation (Pearson's $r = 0.919$, $P < 0.0001$; $n = 119$ cells from 16 mice). **e**, Mean AMPAR/NMDAR ratio for renouncing (R) and persevering (P) mice (ANOVA followed by two-sided t -test: $t_{17} = 3.69$, $*P = 0.002$ and $t_{18} = 4.20$, $*P = 0.0003$ for tdTomato⁺ and tdTomato⁻, respectively; $t_{78} = 6.72$, $*P < 0.0001$, renouncing versus persevering for unidentified neurons, same sample size as in c). **f**, AMPAR/NMDAR ratio for naive mice and for mice yoked to renouncing or persevering mice, with Drd1a–tdTomato identification (12 tdTomato⁺ and 12 tdTomato⁻ cells from, respectively, 4 and 6 naive mice; 6 tdTomato⁺ and 9 tdTomato⁻ cells from 2 mice yoked to renouncing mice and 7 tdTomato⁺ and 8 tdTomato⁻ cells from 2 mice yoked to persevering mice). Data are mean \pm s.e.m. in all panels. See Supplementary Table 1 for complete statistics. **a**, Line drawings modified from Paxinos and Franklin⁴⁴, copyright © 2007.

compared to renouncing mice. In fact, there was a strong correlation between the mean AMPAR/NMDAR ratio and the perseverance (Fig. 4c–e and Extended Data Fig. 7a). In animals yoked to renouncing or persevering mice (that is, mice that only receive the shock, but have never experienced oDASS), the AMPAR/NMDAR ratio was not different from naive mice, demonstrating that the plasticity did not reflect the number of shocks received (Fig. 4f and Extended Data Fig. 7a). The release probability was also higher in persevering mice, as determined by a decrease in the paired-pulse ratio (Extended Data Fig. 7b); moreover, no change in the composition of AMPAR subunits was detected (Extended Data Fig. 6c). Because the ratio of excitatory to inhibitory postsynaptic currents was increased to the same extent as the AMPAR/NMDAR ratio in persevering mice, the OFC to interneuron synapses probably remained unaffected (Extended Data Fig. 7d). Taken together, perseverance was associated with a strengthening of OFC–striatum transmission onto SPNs.

Bidirectional shift in synaptic strength

We next tested whether potentiation of OFC–striatum transmission would lead to perseverance in renouncing mice and, conversely, whether depotentiation in persevering mice would reduce compulsivity.

We found that brief stimulation at 20 Hz of the OFC to dorsal striatum projections was sufficient to potentiate AMPAR excitatory postsynaptic currents (EPSCs) in slices from renouncing ($198 \pm 23\%$) or naive ($180 \pm 20\%$) mice, whereas the long-term potentiation (LTP) was occluded in slices from persevering mice ($111 \pm 15\%$; Fig. 5a). This protocol was then delivered in vivo through optic fibres that targeted OFC terminals in the striatum, and the efficacy was verified ex vivo by measuring an increased AMPAR/NMDAR ratio in slices (Fig. 5b and Extended Data Fig. 8a). In renouncing mice, when applied before the punishment sessions, this procedure led to a significant reduction in the delay to engage in a new sequence and strongly increased the overall oDASS rate (Fig. 5c). This behavioural change was long-lasting and observed even when punishment sessions without OFC–striatum stimulation were added (Fig. 5d). Synaptic potentiation had no effect on the baseline oDASS rate (Extended Data Fig. 8b).

To depotentiate the OFC–striatum transmission, we used two distinct approaches. First, a protocol²⁴ (10 Hz for 5 min) was used that restored presynaptic transmission ex vivo in slices, but that had no effect on behaviour when applied in vivo (Extended Data Fig. 9a–e). Second, low-frequency stimulation (1 Hz for 5 min, typically inducing NMDAR-dependent long-term depression (LTD) that is postsynaptically expressed²⁵) in oDASS mice that expressed Chrimson in the OFC reliably depotentiated synapses in slices from renouncing or naive mice, while yielding very variable results in slices from persevering mice (Fig. 5e). This might be due to ambient dopamine blocking the LTD expression in D1R-SPNs²⁶, which was confirmed when the combination of 1-Hz stimulation and SCH23390 (a D1 receptor (D1R) antagonist) unmasked a synaptic depression in persevering mice and normalized the AMPAR/NMDAR ratio ex vivo (Fig. 5f). The paired-pulse ratio and rectification index remained unchanged (Extended Data Fig. 10a). When applied in vivo before a punished session, the protocol significantly reduced the oDASS rate in a within-control experiment, whereas separate 1-Hz stimulation or SCH23390 application had no effect (Fig. 5g). oDASS was not affected when manipulations were applied before baseline sessions (Extended Data Fig. 10b). Notably, this procedure primarily affected the delay in trial initiation after punishment, similar to the temporal profile that was typically observed in renouncing mice. In contrast to the acute inhibition of OFC–striatum, the effect of LTD on behaviour was still detectable for days, even after cessation of the treatment (Fig. 5h).

Discussion

We found that in a subpopulation of mice that acquired oDASS, the strengthening of OFC–striatum synapses was causally linked to the perseverance of reinforcement despite punishment. Chr2 expression in the VTA was homogeneous and thus did not segregate with behaviour. In addition, the oDASS rate during acquisition and the breakpoint for the progressive ratio schedule were not different between mice that eventually became compulsive and those that renounced oDASS when punished. This is in contrast to previous reports in rats that self-administered cocaine^{27,28}, possibly reflecting differences in the rewarding value of the drug or the training procedure. Thus, the emergence of the two behavioural phenotypes cannot be explained by differences in the intrinsic motivational properties that arise from the stimulation of dopaminergic neurons.

The transient activity of striatum-projecting OFC neurons segregated with the behavioural phenotype, which may constitute the signal for reinforced responding despite punishment. This is consistent with the OFC encoding the expected relative outcome value by updating prior information on reward and/or punishment reception^{19,20,29}. Two scenarios are possible; either the value of the reward becomes excessive or the aversive nature of the punishment is discarded. On one hand, pain perception remains unaffected by oDASS¹⁶; on the other hand,

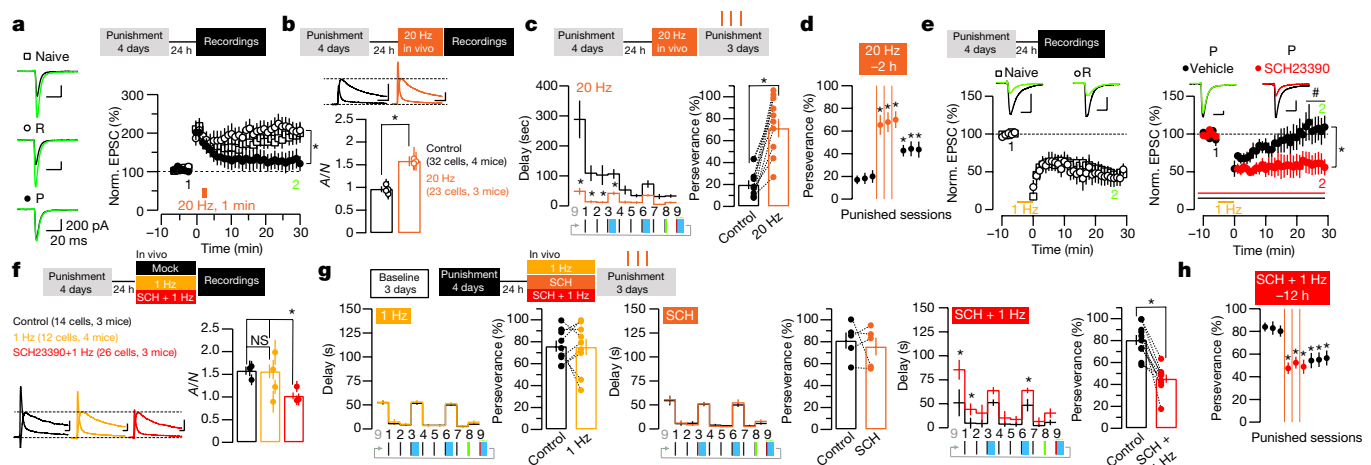


Fig. 5 | Bidirectional modulation of compulsive oDASS. **a**, Average traces for EPSCs recorded immediately before and 30 min after the LTP protocol (20 Hz for 1 min) and group data for normalized (norm.) EPSCs (renouncing and persevering; two-sided t -test: $t_{13} = 3.08$, $*P = 0.009$; 8 and 7 cells from 4 and 3 mice, respectively). Group data are mean \pm s.e.m. of 10 cells from 3 naive mice. **b**, Ex vivo measurements of the AMPAR/NMDAR ratio after in vivo stimulation of OFC–striatum terminals at 20 Hz for 1 min in renouncing mice (control and 20 Hz: $t_{53} = 5.07$, $*P < 0.0001$; $n = 32$ cells from 4 mice and 23 cells from 3 mice, respectively). **c**, Punished sessions in renouncing mice performed 4 h after LTP induction. Delays between lever presses are reduced (ANOVA followed by two-sided t -test: $*P < 0.05$ when comparing delays during punished sessions for control and 20 Hz) and perseverance is increased (two-sided paired t -test: $t_{10} = 6.87$, $*P < 0.0001$; $n = 11$ mice for control and 20 Hz). **d**, During additional punished sessions without renewal of the intervention, perseverance remained higher (ANOVA followed by Dunnett's test: $q_{10} = 5.41$, $*P = 0.0014$; $q_{10} = 5.96$, $*P = 0.0007$; $q_{10} = 5.97$, $*P = 0.0007$; and $q_{10} = 4.31$, $*P = 0.0071$; $q_{10} = 4.19$, $*P = 0.0085$; $q_{10} = 3.64$, $*P = 0.020$; for every punished versus 20 Hz session applied 2 h before and punished versus recovery session, respectively, $n = 11$ mice). **e**, Average traces for EPSCs recorded immediately before and 30 min after the LTD protocol (1 Hz for 5 min) and group data for normalized EPSCs of naive and renouncing mice (left) and persevering mice (right). Right, in slices from persevering mice, LTD is unmasked by bath application of SCH23390 (two-sided t -test comparing persevering and renouncing mice: $t_{13} = 3.25$, $*P = 0.006$; 104.6% for persevering mice ($n = 10$ cells

from 6 mice) versus 48.6% for renouncing mice ($n = 9$ cells from 3 mice); two-sided t -test comparing vehicle versus SCH23390: $t_{14} = 3.07$, $*P = 0.008$; 104.6% for vehicle ($n = 10$ cells from 6 mice) versus 52.6% for SCH23390 in persevering mice ($n = 8$ cells from 2 mice)). **f**, AMPAR/NMDAR ratio was normalized by in vivo stimulation of OFC–striatum terminals with 1 Hz in the presence of SCH23390 in persevering mice (ANOVA followed by two-sided t -test: $t_{24} = 0.34$, $P > 0.99$; $t_{36} = 4.41$, $*P = 0.0002$; and $t_{38} = 4.23$, $*P = 0.0003$; for control versus 1 Hz (14 and 12 cells, respectively); 1 Hz versus 1 Hz with SCH23390 (12 and 26 cells, respectively) and control versus 1 Hz with SCH23390 (14 and 26 cells, respectively)). NS, not significant. **g**, Delay between lever presses in punished sessions 12 h after SCH23390, 1 Hz or 1 Hz with SCH23390 (ANOVA followed by t -test: $*P < 0.05$ when comparing delays during punished sessions for control versus SCH23390, control versus 1 Hz or control versus 1 Hz with SCH23390). Perseverance is reduced in the group of 1 Hz with SCH23390 (two-sided t -test: $t_8 = 0.1$, $P = 0.93$, $n = 9$ mice for control and 1 Hz; $t_4 = 0.6$, $P = 0.60$, $n = 5$ for control versus SCH23390; $t_9 = 6.37$, $*P = 0.0001$, $n = 10$ for control versus 1 Hz with SCH23390). **h**, During additional punished sessions without renewal of the intervention, perseverance reduction remained (ANOVA followed by Dunnett's test: $q_9 = 4.66$, $*P = 0.0054$; $q_9 = 4.43$, $*P = 0.0074$; $q_9 = 5.14$, $*P = 0.0028$; and $q_9 = 3.27$, $*P = 0.040$; $q_9 = 3.61$, $*P = 0.0024$; $q_9 = 3.72$, $*P = 0.0021$ for every punished versus 1 Hz with SCH23390 session applied 12 h before and punished versus recovery session, respectively; $n = 10$ mice). Data are mean \pm s.e.m. See Supplementary Table 1 for complete statistics.

the signal revealed by fibre photometry precedes the initiation of the action and thus cannot reflect the perception of the punishment. In rats, activity of the OFC just before a lever press also correlates with compulsive cocaine self-administration¹⁷, which may be linked to the representation of the expected reward even when there is a risk of punishment. If one accepts that the activity of the OFC increases with goal-directed actions³⁰ then compulsive oDASS may constitute an extreme form of goal-directed behaviour.

The main finding of our study is the identification of the synaptic strengthening of OFC–dorsal striatum projections as a neural mechanism that underlies compulsion. An important question is how does oDASS compare to cocaine exposure. Drug-evoked synaptic plasticity could be mimicked with optogenetic stimulation, starting with the strengthening of excitatory afferents onto dopaminergic neurons in the VTA after the first exposure to an addictive drug^{21,31}. With chronic protocols, oDASS selectively elicits a synaptic potentiation in D1R-SPNs in the nucleus accumbens in every animal¹⁶ that is indistinguishable from the plasticity that is observed after cocaine self-administration^{32–36}, both of which are associated with cue-evoked seeking behaviour. By contrast, the plasticity described here at OFC–striatum synapses was stochastically observed. Differences in the susceptibility rate to compulsive behaviour that was seen in oDASS versus cocaine self-administration (20% for cocaine self-administration³⁷ versus 60% found here) could also reflect different dopamine kinetics and recruited targets (for example, blocking serotonin reuptake) in the two paradigms³⁸.

Another open question is how the plasticity that underlies compulsion is induced. Because the innervation of the striatum by dopaminergic neurons from the VTA is weak, induction is unlikely to be a direct consequence of dopamine release. Moreover, the rules of plasticity dictated by dopamine modulation are not compatible with an induction in both D1R- and D2R-expressing SPNs²⁶. Regardless, the compulsion-associated plasticity in the striatum acts in concert with the involvement of more and more dorsal regions of the striatum as addiction develops³⁹.

Why only a fraction of mice lose control remains to be determined; the emergence of the two groups is even more surprising given the high degree of genetic homogeneity of the mouse line used here (our *Dat^{Cre}* (also known as *Slc6a3^{Cre}*) mice were backcrossed for more than ten generations into the C57BL/6J mouse line), and may reflect a case of stochastic individuality⁴⁰. The emerging circuit model may help to guide molecular investigations while taking into account life experiences. For example, impulsivity has been proposed to be a predictive endophenotype for addiction⁴¹.

The identification of adaptation of a cortical circuit that underlies the late stage of addiction enables a rational refinement of the therapeutic interventions that are currently tested in people with addiction using pharmacology, deep brain stimulation or transcranial magnetic stimulation^{42,43}.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0789-4>

Received: 11 May 2018; Accepted: 13 November 2018;
Published online 19 December 2018.

1. Lüscher, C. & Ungless, M. A. The mechanistic classification of addictive drugs. *PLoS Med.* **3**, e437 (2006).
2. Di Chiara, G. et al. Dopamine and drug addiction: the nucleus accumbens shell connection. *Neuropharmacology* **47**, 227–241 (2004).
3. Keiflin, R. & Janak, P. H. Dopamine Prediction errors in reward learning and addiction: from theory to neural circuitry. *Neuron* **88**, 247–263 (2015).
4. Koob, G. F. Antireward, compulsivity, and addiction: seminal contributions of Dr. Athina Markou to motivational dysregulation in addiction. *Psychopharmacology* **234**, 1315–1332 (2017).
5. Smith, R. J. & Laiks, L. S. Behavioral and neural mechanisms underlying habitual and compulsive drug seeking. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **87**, 11–21 (2018).
6. Vanderschuren, L. J. M. J. & Everitt, B. J. Behavioral and neural mechanisms of compulsive drug seeking. *Eur. J. Pharmacol.* **526**, 77–88 (2005).
7. Volkow, N. D., Koob, G. F. & McLellan, A. T. Neurobiologic advances from the brain disease model of addiction. *N. Engl. J. Med.* **374**, 363–371 (2016).
8. Dalley, J. W., Everitt, B. J. & Robbins, T. W. Impulsivity, compulsivity, and top-down cognitive control. *Neuron* **69**, 680–694 (2011).
9. Yücel, M. et al. A transdiagnostic dimensional approach towards a neuropsychological assessment for addiction: an international Delphi consensus study. *Addiction* <https://doi.org/10.1111/add.14424> (2018).
10. Everitt, B. J. & Robbins, T. W. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nat. Neurosci.* **8**, 1481–1489 (2005).
11. Vandaale, Y. & Janak, P. H. Defining the place of habit in substance use disorders. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **87**, 22–32 (2017).
12. Everitt, B. J. & Robbins, T. W. Drug addiction: updating actions to habits to compulsions ten years on. *Annu. Rev. Psychol.* **67**, 23–50 (2016).
13. McCracken, C. B. & Grace, A. A. Persistent cocaine-induced reversal learning deficits are associated with altered limbic cortico-striatal local field potential synchronization. *J. Neurosci.* **33**, 17469–17482 (2013).
14. Chen, B. T. et al. Rescuing cocaine-induced prefrontal cortex hypoactivity prevents compulsive cocaine seeking. *Nature* **496**, 359–362 (2013).
15. Jonkman, S., Pelloux, Y. & Everitt, B. J. Differential roles of the dorsolateral and midlateral striatum in punished cocaine seeking. *J. Neurosci.* **32**, 4645–4650 (2012).
16. Pascoli, V. et al. Sufficiency of mesolimbic dopamine neuron stimulation for the progression to addiction. *Neuron* **88**, 1054–1066 (2015).
17. Guillem, K. & Ahmed, S. H. Preference for cocaine is represented in the orbitofrontal cortex by an increased proportion of cocaine use-coding neurons. *Cereb. Cortex* **28**, 819–832 (2018).
18. Lucantonio, F., Stalnaker, T. A., Shaham, Y., Niv, Y. & Schoenbaum, G. The impact of orbitofrontal dysfunction on cocaine addiction. *Nat. Neurosci.* **15**, 358–366 (2012).
19. Lucantonio, F. et al. Effects of prior cocaine versus morphine or heroin self-administration on extinction learning driven by overexpectation versus omission of reward. *Biol. Psychiatry* **77**, 912–920 (2015).
20. Schoenbaum, G., Chang, C. Y., Lucantonio, F. & Takahashi, Y. K. Thinking outside the box: orbitofrontal cortex, imagination, and how we can treat addiction. *Neuropsychopharmacology* **41**, 2966–2976 (2016).
21. Brown, M. T. C., Korn, C. & Lüscher, C. Mimicking synaptic effects of addictive drugs with selective dopamine neuron stimulation. *Channels* **5**, 461–463 (2011).
22. Sciamanna, G., Ponterio, G., Mandolesi, G., Bonsi, P. & Pisani, A. Optogenetic stimulation reveals distinct modulatory properties of thalamostriatal vs corticostriatal glutamatergic inputs to fast-spiking interneurons. *Sci. Rep.* **5**, 16742 (2015).
23. Gerfen, C. R. et al. D1 and D2 dopamine receptor-regulated gene expression of striatonigral and striatopallidal neurons. *Science* **250**, 1429–1432 (1990).
24. Grueter, B. A., Brasnjic, G. & Malenka, R. C. Postsynaptic TRPV1 triggers cell type-specific long-term depression in the nucleus accumbens. *Nat. Neurosci.* **13**, 1519–1525 (2010).
25. Pascoli, V., Turiault, M. & Lüscher, C. Reversal of cocaine-evoked synaptic potentiation resets drug-induced adaptive behaviour. *Nature* **481**, 71–75 (2012).
26. Shen, W., Flajolet, M., Greengard, P. & Surmeier, D. J. Dichotomous dopaminergic control of striatal synaptic plasticity. *Science* **321**, 848–851 (2008).
27. Pelloux, Y., Everitt, B. J. & Dickinson, A. Compulsive drug seeking by rats under punishment: effects of drug taking history. *Psychopharmacology* **194**, 127–137 (2007).
28. Kasanetz, F. et al. Transition to addiction is associated with a persistent impairment in synaptic plasticity. *Science* **328**, 1709–1712 (2010).
29. Lucantonio, F. et al. Orbitofrontal activation restores insight lost after cocaine use. *Nat. Neurosci.* **17**, 1092–1099 (2014).
30. Padoa-Schioppa, C. & Conen, K. E. Orbitofrontal cortex: a neural circuit for economic decisions. *Neuron* **96**, 736–754 (2017).
31. Ungless, M. A., Whistler, J. L., Malenka, R. C. & Bonci, A. Single cocaine exposure *in vivo* induces long-term potentiation in dopamine neurons. *Nature* **411**, 583–587 (2001).
32. Pascoli, V. et al. Contrasting forms of cocaine-evoked plasticity control components of relapse. *Nature* **509**, 459–464 (2014).
33. Terrier, J., Lüscher, C. & Pascoli, V. Cell-type specific insertion of GluA2-lacking AMPARs with cocaine exposure leading to sensitization, cue-induced seeking, and incubation of craving. *Neuropsychopharmacology* **41**, 1779–1789 (2016).
34. Hearing, M., Graziane, N., Dong, Y. & Thomas, M. J. Opioid and psychostimulant plasticity: targeting overlap in nucleus accumbens glutamate signaling. *Trends Pharmacol. Sci.* **39**, 276–294 (2018).
35. Lüscher, C. The emergence of a circuit model for addiction. *Annu. Rev. Neurosci.* **39**, 257–276 (2016).
36. Wolf, M. E. Synaptic mechanisms underlying persistent cocaine craving. *Nat. Rev. Neurosci.* **17**, 351–365 (2016).
37. Deroche-Gamonet, V., Belin, D. & Piazza, P. V. Evidence for addiction-like behavior in the rat. *Science* **305**, 1014–1017 (2004).
38. Pelloux, Y., Dilleen, R., Economidou, D., Theobald, D. & Everitt, B. J. Reduced forebrain serotonin transmission is causally involved in the development of compulsive cocaine seeking in rats. *Neuropsychopharmacology* **37**, 2505–2514 (2012).
39. Belin, D. & Everitt, B. J. Cocaine seeking habits depend upon dopamine-dependent serial connectivity linking the ventral with the dorsal striatum. *Neuron* **57**, 432–441 (2008).
40. Honegger, K. & de Bivort, B. Stochasticity, individuality and behavior. *Curr. Biol.* **28**, R8–R12 (2018).
41. Belin, D., Mar, A. C., Dalley, J. W., Robbins, T. W. & Everitt, B. J. High impulsivity predicts the switch to compulsive cocaine-taking. *Science* **320**, 1352–1355 (2008).
42. Diana, M. et al. Rehabilitating the addicted brain with transcranial magnetic stimulation. *Nat. Rev. Neurosci.* **18**, 685–693 (2017).
43. Coles, A. S., Kozak, K. & George, T. P. A review of brain stimulation methods to treat substance use disorders. *Am. J. Addict.* **27**, 71–91 (2018).
44. Paxinos, G. & Franklin, K. B. J. *The Mouse Brain in Stereotaxic Coordinates* (Academic, New York, 2007).

Acknowledgements We thank E. C. O'Connor for discussion and comments on the manuscript; C. Gerfen for providing Cre-mouse lines through the MMRC repository. This study was financed by a grant from the Swiss National Science Foundation (Ambizione grant to P.V. and core grant to C.L.), the National Center of Competence in Research (NCCR) SYNAPSY-The Synaptic Bases of Mental Diseases, and an advanced grant from the European Research Council (MeSSI).

Reviewer information Nature thanks J. P. Britt and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions V.P. conceived the experiments and performed patch recordings and behavioural experiments. A.H. did surgeries for viral infection, behavioural experiments and *in vivo* recordings. R.A. carried out the retrograde tracing with the rabies strategy. R.V.Z. implanted the photometry and carried out analyses and recordings. M.L. and M.H. carried out patch recordings. J.F. carried out clustering analysis. C.L. conceptualized and supervised the study, and prepared the manuscript with the help of all authors.

Competing interests C.L. is a member of the following scientific advisory boards: Stalicia SA; Phénix Foundation; International research in paraplegia (IRP) Foundation, Geneva.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0789-4>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0789-4>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to C.L.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Animals. Mice (age 8–24 weeks) were heterozygous BAC-transgenic mice in which the Cre-recombinase expression was under the control of the regulatory elements of the dopamine transporter gene (DAT-Cre mice⁴⁵). *Dat^{cre}* mice were originally provided by G. Schutz and only heterozygous mice were used for experiments. DAT B6.SJL-Slc6a3tm1.1(cre)Bkmm/J (also known as DAT-IRES-Cre) mice crossed with mice in which tomato expression was driven by the D1R (*Drd1a^{tdTomato}* from Jackson Laboratories) gene regulatory element were also used. Tg(*Drd1-cre*)120Mxu, Tg(*Drd2-cre*)ER43Gsat and 129P2-Pvalb^{tm1(cre)}Arbr/J mice from Charles River were used for tracing studies. Weights and genders were distributed homogeneously among the groups. Transgenic mice had been backcrossed into the C57BL/6 line for a minimum of four generations. Mice were single-housed after surgery. All animals were kept in a temperature- and humidity-controlled environment with a 12-h light/12-h dark cycle (lights on at 07:00). All procedures were approved by the Institutional Animal Care and Use Committee of the University of Geneva and by the animal welfare committee of the Cantonal of Geneva, in accordance with Swiss law.

Stereotaxic injections. AAV5-EF1a-DIO-ChR2(H134R)-eYFP or AAV8-hSyn-Flex-ChrimsonR-tdTomato produced at the University of North Carolina (UNC Vector Core Facility) were injected into the VTA of 5- to 6-week-old mice. Anaesthesia was induced at 5% and maintained at 2.5% isoflurane (w/v) (Baxter) during surgery. The mouse was placed in a stereotaxic frame (Angle One) and craniotomies were performed using stereotaxic coordinates (for VTA: anterior–posterior (AP) –3.3; medial–lateral (ML) –0.9 with a 10° angle; dorsal–ventral (DV) –4.3). Injections of virus (0.5 µl) used graduated pipettes (Drummond Scientific), broken back to a tip diameter of 10–15 mm, at an infusion rate of 0.05 µl min^{–1}. Following the same procedure, AAV8-hSyn-ChrimsonR-tdTomato (UNC), AAV8-hSyn-Chrimson-GFP (Duke University), AAV5-CamKII-eArchT3.0-eYFP (UNC) or AAV-DJ-CamKII-GCaMP6m (Stanford University) was injected bilaterally in the OFC (AP +2.6; ML ±1.60; DV –1.8). During the same surgical procedure, a unique chronically indwelling optic fibre cannula was implanted above the VTA using the exact same coordinates as for the injection except for DV coordinate, which was reduced to 4.2. Three screws were fixed into the skull to support the implant, which was further secured with dental cement. The first behavioural session typically occurred 10–14 days after surgery to allow sufficient expression of ChR2 or Chrimson. A photometry fibre was implanted unilaterally in the striatum (AP +0.8; ML 1.75; DV –2.4), fibres for eArchT3.0 stimulation were placed in each OFC (AP +2.6; ML ±1.60; DV –1.6) and fibres for terminals stimulation in the striatum were implanted bilaterally (AP +0.8; ML ±1.75; DV –2.4). In D1R-Cre, D2R-Cre and PV-Cre mice, AAV5-EF1a-Flex-TVA-mCherry, AAV8-CA-Flex-RG and EnvA G-deleted Rabies-eGFP from Salk Institute were injected in the striatum (AP +0.6; ML +1.8; DV –3.2) a week before the pseudotyped rabies virus.

Optogenetic self-stimulation apparatus. Mice infected with ChR2 (or Chrimson) in the VTA were placed during the light phase in operant chambers (ENV-307A-CT, Med Associates) situated in a sound-attenuating box (Med Associates). The optic fibre of the mouse was connected to DPSS blue- or orange-light lasers (CNI-473-140-10-LED-TTL1-MM200FC; CNI-593-200-10-LED-TTL1-MM200FC; Laser 2000) via a FC/PC fibre cable (M72L02; Thorlabs) and a simple rotary joint (FRJ-FC-FC; Doric Lenses) allowing free movement during operant behaviour. Power at the exit of the patch cord was set to 15 ± 1 mW. Two retractable levers were present on one wall of the chamber and a cue light was located above each lever. A cage light was present in each chamber. A rack mount interface cabinet (SG-6010A, Med Associates) containing a programmable constant current shocker (ENV-413, Med Associates) connected to a quick disconnect grid harness (ENV-307A-QD, Med Associates) was used to provide foot shocks during punishment sessions. The apparatus was controlled and data captured using MED-PC IV software (Med Associates). For acute stimulation in the OFC or at terminals of OFC fibres in the striatum, a double rotary joint (FRJ_1x2i_FC-2FC, Doric Lenses) was used to connect cables to each hemisphere. For orange laser stimulation of eArchT3.0 or Chrimson, a shutter (CMSA-SR475_FC, Doric Lenses) was used to avoid variation in intensities during laser warm-up.

Self-stimulation acquisition and progressive ratio. *Dat^{cre}* mice learned to self-stimulate dopaminergic neurons in the VTA infected with AAV5-DIO-ChR2-eYFP (oDASS) for 12 consecutive days. Each of the 12 acquisition sessions lasted 120 min or until the mouse reached 80 optogenetic stimulations, whichever came first. In each oDASS session, mice could respond on the active lever, resulting in VTA stimulation. Responding on the inactive lever had no consequences. During the first 4 sessions, a single press on the active lever (termed fixed-ratio one or FR1) resulted in a 10-s illumination of a cue light (pulses of 1 s at 1 Hz). After a delay of 5 s, onset of a 15-s laser stimulation (473 nm) composed of 30 bursts separated by 250 ms (each burst consisted of 5 laser pulses of 4-ms pulse width at 20 Hz). A 20-s time-out followed the rewarded lever press, during which lever presses had no consequences but were still recorded. Next, a FR2 (sessions 5–8) and a FR3 (sessions 9–12) were introduced.

For measurements of motivation, mice did a single progressive ratio session between acquisition sessions 11 and 12 that lasted for a maximum of 4 h. The breakpoint was considered to be the last reached reinforced schedule after 4 h or after 40 min had elapsed since the last reinforced schedule. The reinforced schedules were the following: 1, 3, 5, 8, 12, 16, 22, 29, 38, 50, 65, 84, 108, 139, 178, 228, 291, 371, 473, 603, 767, 977, 1,243 and 1,582. The total number of active lever presses instead of the breakpoint was represented in Fig. 1 to better visualize individual performance.

Punishment sessions. After acquisition, mice underwent 3 additional sessions with a reduced cut-off (maximum 40 laser stimulations or 60 min, whichever came first). These sessions served as a baseline before starting the punishment sessions. Punishment sessions occurred in the exactly same conditions as for baseline sessions, except every third laser stimulation event was preceded by a foot shock (500 ms, 0.25 mA) starting immediately after the completion of the FR3 (5 s before the onset of the laser stimulation). Impending punishment was announced by illumination of the home cage light for 2-s after the second press of the FR3. Indeed, a new cue (cage light) predicting the oncoming shock was paired with the second lever press of the FR3 schedule, directly preceding the shock-coupled press.

Fibre photometry recordings. Fibre photometry recordings were performed during baseline and punished sessions in oDASS mice. A batch of DAT-Cre animals infected with Chrimson (AAV8-hSyn-Flex-ChrimsonR-tdTomato produced at UNC) in the VTA and with GCaMP6m (AAV-DJ-CamKII-GCaMP6m, Stanford University) in the OFC were used for recordings of the activity of terminals in the striatum during oDASS. Mice were recorded for 20–40 min per session to minimize bleaching. OFC–striatum terminals were illuminated with blue (470 nm wavelength, M470F3, Thorlabs) and violet (405 nm wavelength, M405FP1, Thorlabs) filtered excitation LED lights, that were sinusoidally modulated at 211 and 531 Hz. Green emission light (500–550 nm) was collected through the same fibre that was used for excitation and passed onto a photoreceiver (Newport 2151, Doric Lenses). Pre-amplified signals were then demodulated by a real-time signal processor (RZ5P, Tucker Davis Systems) to determine contributions from 470 nm and 405 nm excitation sources⁴⁶. TTL signals of the relevant stimuli were directly sent from the operant chamber to the signal processor. Analysis was performed offline in MATLAB. To calculate $\Delta F/F_0$, a linear fit was applied to the 405-nm control signal to align it to the 470-nm signal. This fitted 405-nm signal was used as F_0 in standard $\Delta F/F_0$ normalization $((F(t) - F_0(t))/F_0(t))$. Averaged peristimulus activity traces were then constructed for which the mean baseline fluorescence (–1.5 to –0.5 s before the relevant event) was subtracted from the trace.

Fibre photometry data. To identify active lever-press (ALP)-related $\Delta F/F_0$ signal modulations over time, we examined the signal across mice (renouncing versus persevering) aligned to the ALP, using a time bin of 10 ms, following a published protocol^{47,48}. Baseline intervals was taken from –1.5 to –0.5 s and the event from –0.5 to 1.5 s, followed by the calculation of the average of the mean ± s.d. $\Delta F/F_0$ values of the baseline. We then checked bin-by-bin (100 ms) for a threshold of 1.65 s.d. (95% confidence interval) and significant modulation was found if 20 consecutive bins passed the threshold test.

To compare trials from the same animal, we used a permutation test^{49,50}. In brief, we extracted all trials aligned on the ALP in the interval –1.5 to 1.5 s for the two conditions. We then collected all trials of the baseline and punishment sessions and randomly drew from this combined set as many trials as in the baseline session (subset 1) while the remaining trials were placed in subset 2. This random partition was repeated 1,000 times to compute the $\Delta F/F_0$ mean and s.d. for these two permutation subsets across the trials at every time point (bin size = 10 ms). Significance was achieved if 20 consecutive bins met the threshold of 3.29 s.d. (99% confidence interval).

Acute inhibition of the OFC during oDASS. Acute inhibition of the OFC transmission during oDASS sessions was performed in mice infected with AAV5-CamKII-eArchT3.0-eYFP in the OFC and ontogenetic fibres were bilaterally implanted in the OFC. Acute inhibition started at a specific epoch of behaviour and lasted for a maximum of 90 s in animals that were infected in the OFC with eArchT3.0 in the OFC. Amber laser light started at different epochs of behaviour: (1) immediately after the active lever press that triggered the punishment-predictive cue (home cage light) until the next press or for a maximum of 90 s, (2) after oDASS of the punished trial until the next press or for a maximum of 90 s, and (3) immediately after every oDASS until the next press or for a maximum of 90 s. Stimulation consisted of continuous laser activation for 6 s followed by a 2-s time-out period, repeated until the mouse initiated the next epoch or for a maximum of 90 s. Acute inhibition before the initiation of a next trial was also tested during baseline sessions.

Slice electrophysiology. Whole-cell patch-clamp recordings of striatal neurons were performed 24 h after the last punished session, after acquisition or in slices from naive mice. The AMPAR/NMDAR ratio was calculated at +40 mV with the AMPAR component pharmacologically isolated using D-AP5 (50 µM) and the NMDAR EPSC component was determined by subtraction. Currents were evoked

with optogenetic stimulation on slices of OFC terminals infected with Chrimson. When using BAC transgenic (Drd1a-tdTomato) mice crossed with *Dat^{cre}* mice, the OFC was infected with AAV8-hSyn-Chrimson-eYFP and tdTomato⁺ or tdTomato⁻ cells were filled with biocytin and identified post hoc on confocal images.

Coronal 230- μ m slices of mouse brain were prepared in cooled artificial cerebrospinal fluid containing (in mM): NaCl 119, KCl 2.5, MgCl₂ 1.3, CaCl₂ 2.5, Na₂HPO₄ 1.0, NaHCO₃ 26.2 and glucose 11, bubbled with 95% O₂ and 5% CO₂. Slices were kept at 32–34 °C in a recording chamber superfused with 2.5 ml min⁻¹ artificial cerebrospinal fluid.

Ex vivo synaptic properties of the striatum. Visualized whole-cell patch-clamp recording techniques were used to measure synaptic responses to optogenetic stimulation of OFC terminals. In some experiments, D1R- and D2R-SPNs of the striatum were identified by the presence of the tdTomato in BAC transgenic mice by using a fluorescence microscope (Olympus BX50WI, fluorescent light U-RFL-T) and confirmed on confocal images of the recorded neuron filled with biocytin (Sigma, B4261) and stained with streptavidin–Cy5 (Invitrogen, 434316). The holding potential was –70 mV and the access resistance was monitored by a hyperpolarizing step of –14 mV. The liquid junction potential was small (–3 mV), and therefore traces were not corrected. Experiments were discarded if the access resistance varied by more than 20%. Currents were amplified (Multiclamp 700B, Axon Instruments), filtered at 5 kHz and digitized at 20 kHz (National Instruments Board PCI-MIO-16E4, Igor, Wave Metrics). For recordings of optogenetically evoked EPSCs, the internal solution contained (in mM): CsCl 130, NaCl 4, creatine phosphate 5, MgCl₂ 2, Na₂ATP 2, Na₃GTP 0.6, EGTA 1.1, HEPES 5 and spermine 0.1. QX-314 (5 mM) was added to the solution to prevent action currents. Synaptic currents were evoked by short light pulses (4 ms) at 0.1 Hz through an LED (M590L3-C1, Thorlabs) placed through the objective above the tissue. To isolate AMPAR-evoked EPSCs the NMDA antagonist D-2-amino-5-phosphonopivalic acid (D-AP5, 50 μ M) was applied to the bath. The NMDAR component was calculated as the difference between the EPSCs measured in the absence and in the presence of D-AP5. The AMPAR/NMDAR ratio was calculated by dividing the peak amplitudes. The AMPAR/NMDAR ratio was also calculated by taking the NMDAR EPSC component 20 ms after the peak of the EPSCs recorded at +40 mV without pharmacological isolation. The rectification index of AMPAR was calculated as the ratio of the chord conductance calculated at negative potential divided by chord conductance at positive potential. The paired-pulse ratio (PPR) was measured during the first 3 min of the recordings by delivering 2 pulses of 4 ms with a 76-ms interval. Examples traces are averages of 10–15 sweeps. All experiments were performed in the presence of picrotoxin (100 μ M).

In vitro synaptic plasticity. Low-frequency stimulation (1 or 10 Hz for 5 min) was applied with 4-ms light pulses and the magnitude of LTD was determined by comparing average EPSCs that were recorded 20–30 min after induction to EPSCs recorded immediately before induction. These experiments were conducted with bath application of the corresponding vehicle, with the D1R-antagonist SCH23390 (10 μ M) or with the NMDA use-dependent channel blocker MK801 (10 μ M) and the mGluR5-positive allosteric modulator (PAM, CBPPB 30 μ M). For LTD on slices, the internal solution contained (in mM) CsCl 130, NaCl 4, creatine phosphate 5, MgCl₂ 2, Na₂ATP 2, Na₃GTP 0.6, EGTA 1.1, HEPES 5, QX-314 5 and spermine 0.1. For induction of a LTP Chrimson expressed in OFC terminals was stimulated at 20 Hz for 1 min. For LTP experiments, the following internal solution was used (in mM): potassium gluconate 140, MgCl₂ 2, KCl 5, Na₂ATP 4, Na₃GTP 0.3, creatine phosphate 10, HEPES 10 and EGTA 0.2. For validation of terminal inhibition with 3 Hz stimulation for 1 min the internal solution was (in mM): CsCl 130, NaCl 4, creatine phosphate 5, MgCl₂ 2, Na₂ATP 2, Na₃GTP 0.6, EGTA 1.1, HEPES 5, QX-314 5 and spermine 0.1. EPSCs were evoked at 0.1 Hz before and after the protocol (3-Hz stimulation). All experiments were performed in the presence of picrotoxin (100 μ M).

Feed-forward inhibition. For recordings of transmission, EPSCs and inhibitory postsynaptic currents (IPSCs) from OFC to SPNs of the striatum, the internal solution contained (in mM): CsCH₃SO₄ 128, NaCl 20, CaCl₂ 0.3, MgCl₂ 1, Na₂ATP 2, Na₃GTP 0.3, EGTA 1 and HEPES 10. The holding potential was –70 mV for EPSC recordings (reversal potential for GABA (γ -aminobutyric acid)) and 0 mV for IPSC recordings (reversal potential for AMPA). For pharmacological validation, picrotoxin (100 μ M) or NBQX (20 μ M) was added to the bath perfusion during the recordings. The ratio of excitatory to inhibitory transmission was measured as the ratio of the charge transfer obtained at –70 mV and 0 mV for 5 light pulses and was determined at different frequencies (5, 10, 20 and 40 Hz). Charge transfer is determined as the sum of area under the curves for EPSCs or IPSCs.

eArchT3.0 validation. For recordings of OFC pyramidal neurons infected with eArchT3.0 the internal solution contained (in mM): potassium gluconate 130, MgCl₂ 4, Na₂ATP 3.4, Na₃GTP 0.1, creatine phosphate 10, HEPES 5 and EGTA 1.1. Firing was triggered by a current step (1-s duration, 200-pA step) with or without stimulation of eArchT3.0 with the amber LED. No clamp was imposed

and the cell was discarded if the resting membrane potential varied by more than 10%.

In vivo plasticity with optogenetic stimulation protocols and pharmacology.

Optogenetic protocols were applied once in vivo, through bilaterally implanted optical fibres targeting the striatum 2–24 h before a punished session, or before animals were killed for ex vivo recordings. DPSS orange-light lasers (CNI-593-200-10-LED-TTL1-MM200FC, Laser 2000) connected to the indwelling optic fibre via customized patch cords (M72L02, Thorlabs) and a double rotary joint (FRJ_1x2i_FC-2FC, Doric Lenses) allowed mice to move freely during stimulation. The laser was triggered to deliver 4-ms pulses at 1 Hz or 20 Hz for 5 or 1 min, respectively. Optogenetic stimulation was applied in the home cage 4–24 h before the punished session or before the animals were killed for ex vivo electrophysiology recordings. Protocols were also tested before non-punished sessions. SCH23390 (0.3 mg kg⁻¹, 0.1% DMSO; Tocris, 0925) was given intraperitoneally (10 ml kg⁻¹), 20 min before the optogenetic stimulation protocol. CDPBB (30 mg kg⁻¹, 10% Tween 80 (Tocris, 3235) and (+)-MK801 (0.3 mg kg⁻¹ (Tocris, 0924)) was given intraperitoneally (10 ml kg⁻¹), 50 and 20 min before optogenetic stimulation at 10 Hz. Different batches of mice received either the pharmacology or the optogenetic stimulation protocols before testing for perseverance in punished sessions.

Tissue preparation for imaging. Mice were anaesthetized with pentobarbital (300 mg kg⁻¹, intraperitoneally, Sanofi-Aventis) and transcardially perfused with 4% (w/v) paraformaldehyde in PBS (pH 7.5). Brains were post-fixed overnight in the same solution and stored at 4 °C. Coronal or parasagittal oblique sections (70- μ m thick) were cut with a vibratome (Leica), stained with Hoechst (Sigma-Aldrich) and mounted with Mowiol (Sigma-Aldrich). Full images of brain slices were obtained with a Zeiss Axioscan Z1 system equipped with a Plan-Apochromat 10 \times /0.45 NA objective, together with filters for 4',6-diamidino-2-phenylindole (DAPI) (emission band-pass filter: 445/50 nm), enhanced green fluorescent protein (eGFP) (emission band-pass filter: 525/50 nm), cyanine 3 (Cy3) (emission band-pass filter: 605/70 nm) and cyanine 5 (Cy5) (emission band-pass filter: 690/50 nm). Images from VTA, OFC and striatum were obtained using sequential laser scanning confocal microscopy (Zeiss LSM700). Photomicrographs were obtained with the following band-pass and long-pass filter settings: UV excitation (band-pass filter: 365/12 nm), GFP (band-pass filter: 450–490 nm), Cy3 (band-pass filter: 546/12 nm) and Cy5 (band-pass filter: 546/12 nm). For biocytin (Sigma-Aldrich, B4261) staining, streptavidin–Cy5 (Invitrogen 434316) was used, high-magnification images were obtained and a z-stack was made. For immunohistochemistry, the following primary antibody (rabbit polyclonal anti-tyrosine hydroxylase, Millipore AB152, lot 2722866, diluted 1:500) and the secondary antibody (donkey anti-rabbit Cy3, Millipore AP182C, lot 2397069, diluted 1:500) were used.

Clarity. C57BL/6J mice infected with AAV8-Syn-Flex-ChrimsonR-tdt in the OFC or D1R-Cre mice infected with AAV8-hSyn-Flex-TVA-P2A-GFP, AAV8-CA-Flex-RG and EnvA G-deleted Rabies-mCherry-RbE (from Salk Institute) in the striatum were anaesthetized with pentobarbital (300 mg kg⁻¹, intraperitoneally, Sanofi-Aventis) and transcardially perfused with 4% (w/v) paraformaldehyde in PBS (pH 7.5).

Brains were extracted, immersed for 12 h in 4% PFA, rinsed in PBS and immersed for 3 days in the hydrogel monomer solution consisting of 4% acrylamide, 0.25% VA044 Wako thermal initiator. Tubes were flushed with nitrogen gas and tissues were polymerized in a 37 °C water bath for 3 h. Active clearing was achieved using the X-Clarity Tissue clearing system (Logos Biosystems). Fluorescence imaging of CLARITY samples in Histodenz (Sigma-Aldrich, D22158) was performed using a light-sheet fluorescence microscope (Carl Zeiss LSM Z.1) with a Fluor 4 \times objective lens. Images were reconstructed in 3D using TeraStitcher and Imaris software.

Clustering method. Clustering methods allowed identification of renouncing and persevering mice. Clustering was performed on the entire set of behavioural data, namely the number of active and inactive lever presses, the time of the last laser stimulation, the time until the end of the session and finally the delays between the active lever presses (restricted to the third and fourth punished sessions displaying a clear bimodal distribution in Fig. 1c). Prior to the clustering, we applied a nonlinear dimension reduction algorithm (*t*-distributed stochastic neighbor embedding⁵⁰) to end up with two relevant projected variables. We then used a hierarchical clustering algorithm (with the Minkowski metric and average linkage method, using the MATLAB 'pdist', 'linkage' and 'cluster' functions) to divide the mice into two clusters based on their overall behavioural similarities (see Extended Data Fig. 1a). The mean silhouette value (computed with the MATLAB 'silhouette' function) was equal to 0.95, indicating that the clustering solution is appropriate. Mapping the perseverance onto these two clusters supported our initial separation based on visual inspection (see). To construct the ellipse around a cluster, we first fixed the centre as the mean position of the cluster. We then computed from these coordinates a covariance matrix, which was rescaled to ensure that at the end the ellipse enclosed around 98% of the points in the cluster. We got the coordinates of

the ellipse by applying the eigenvectors and eigenvalues of the covariance matrix to the unit circle (that is, for the rotation and scaling), which was finally translated by adding the mean position of the cluster initially computed.

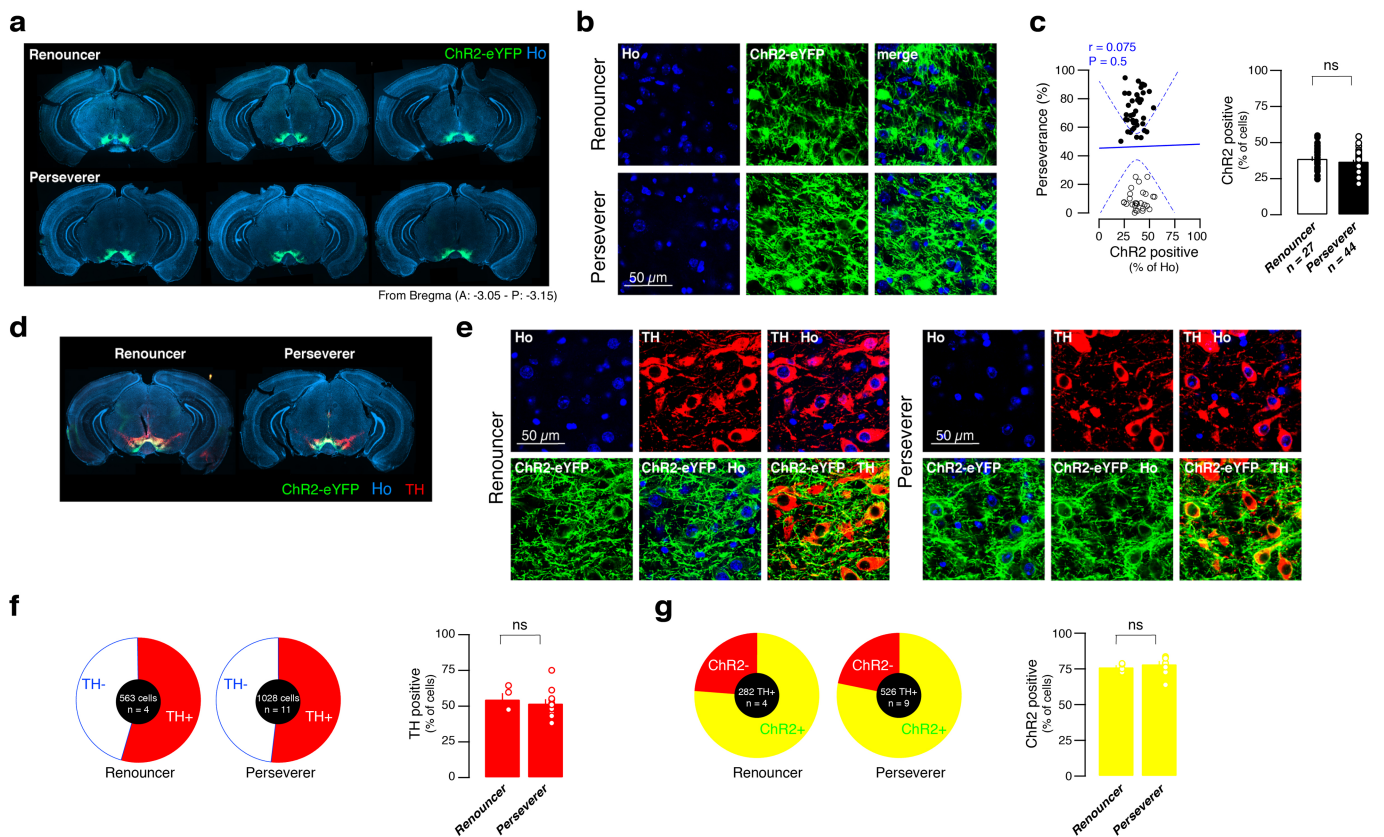
Statistics. Samples sizes were predetermined using a power and sample size calculator. The experiments were not randomized. The investigators were blinded to genotypes during experiments and outcome assessment. Multiple comparisons were first subject to mixed-factor ANOVA defining both between- and/or within-group factors. For comparisons in which significant main effects or interaction terms were found ($P < 0.05$), further comparisons were made by a two-tailed Student's *t*-test with Bonferroni corrections applied when appropriate (that is, the level of significance was 0.05 adjusted by the number of comparisons). Single comparisons of between- or within-group measures were made by two-tailed non-paired or paired Student's *t*-test, respectively.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

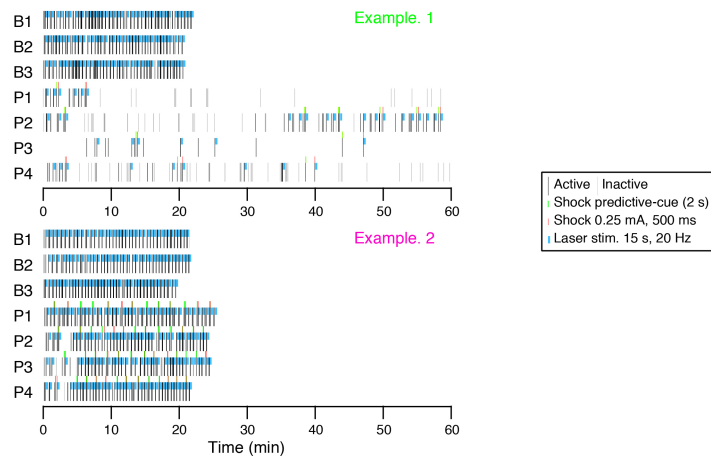
The dataset is available from <https://doi.org/10.5281/zenodo.1474531>.

45. Turiault, M. et al. Analysis of dopamine transporter gene expression pattern — generation of DAT-iCre transgenic mice. *FEBS J.* **274**, 3568–3577 (2007).
46. Lerner, T. N. et al. Intact-brain analyses reveal distinct information carried by SNc dopamine subcircuits. *Cell* **162**, 635–647 (2015).
47. da Silva, J. A., Tecuapetla, F., Paixão, V. & Costa, R. M. Dopamine neuron activity before action initiation gates and invigorates future movements. *Nature* **554**, 244–248 (2018).
48. Li, Y. et al. Serotonin neurons in the dorsal raphe nucleus encode reward signals. *Nat. Commun.* **7**, 10503 (2016).
49. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164**, 177–190 (2007).
50. Van Der Maaten, L. & Hinton, G. H. Visualizing data using *t*-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).



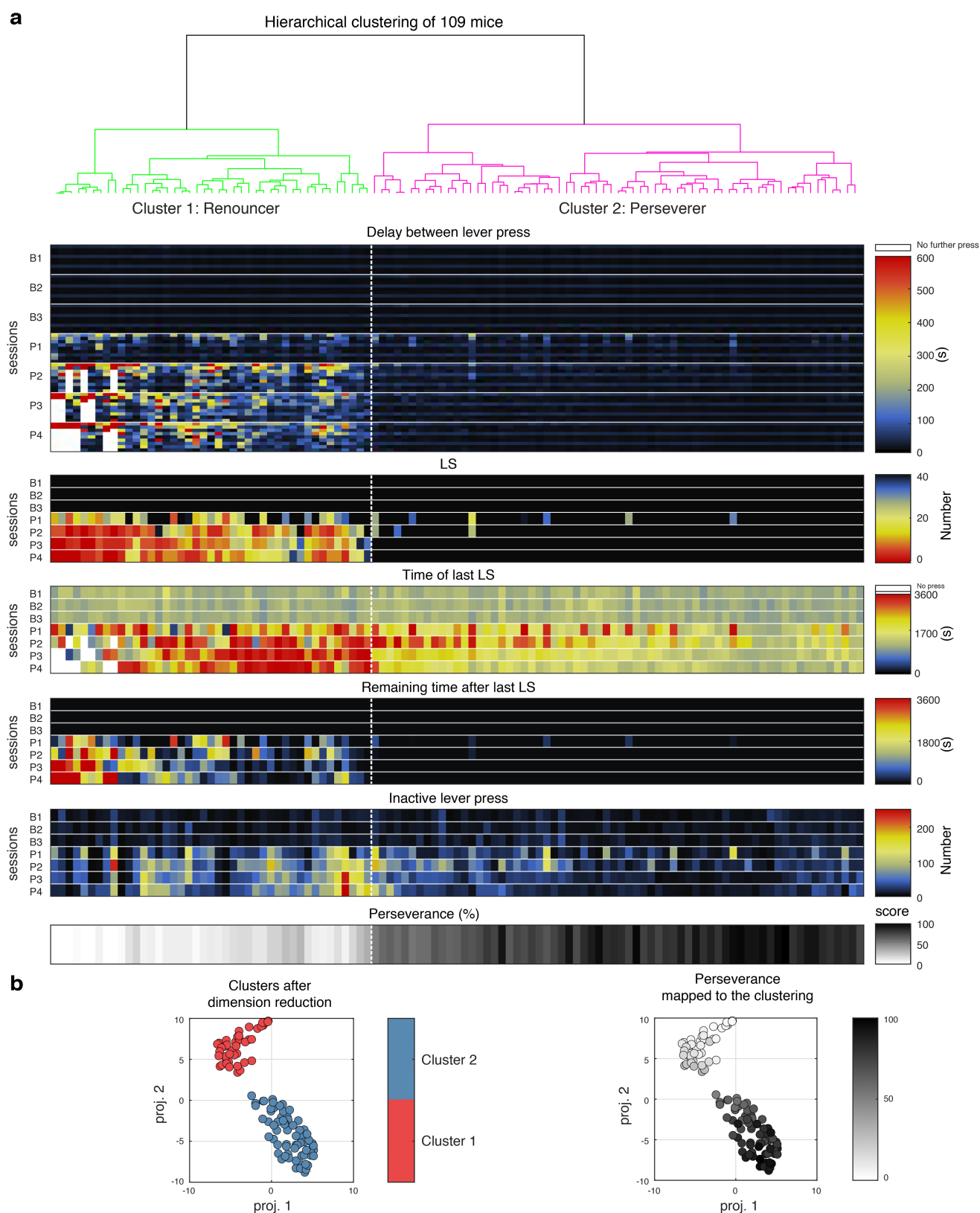
Extended Data Fig. 1 | No correlation between perseverance and VTA infection. **a**, Serial coronal sections of a renouncing and a persevering mouse centred on the VTA infected with AAV5-EF1a-DIO-ChR2-eYFP and nuclear Hoechst staining (Ho). We infected 109 mice and took coronal images of brains from 71 mice. **b**, High-magnification images of VTA. **c**, oDASS perseverance as a function of the infection rate and group data ($n = 71$ mice, Pearson's $r = 0.075$). Infection rate was determined as the number of ChR2-eYFP-positive cells normalized to the total number of cells based on Hoechst staining. Note that mice from which sagittal

sections were obtained are not included in this quantification. **d**, Coronal sections of a renouncing and a persevering mouse that show VTA as above. Sections were additionally stained for tyrosine hydroxylase (TH) using a Cy3-conjugated secondary antibody. Staining was performed in slices from 15 mice. **e**, High-magnification images of VTA. **f**, Quantification of TH-positive neurons in the VTA from a subset of renouncing and persevering mice. **g**, Percentage of TH neurons infected with ChR2. Data are mean \pm s.e.m. See Supplementary Table 1 for complete statistics.



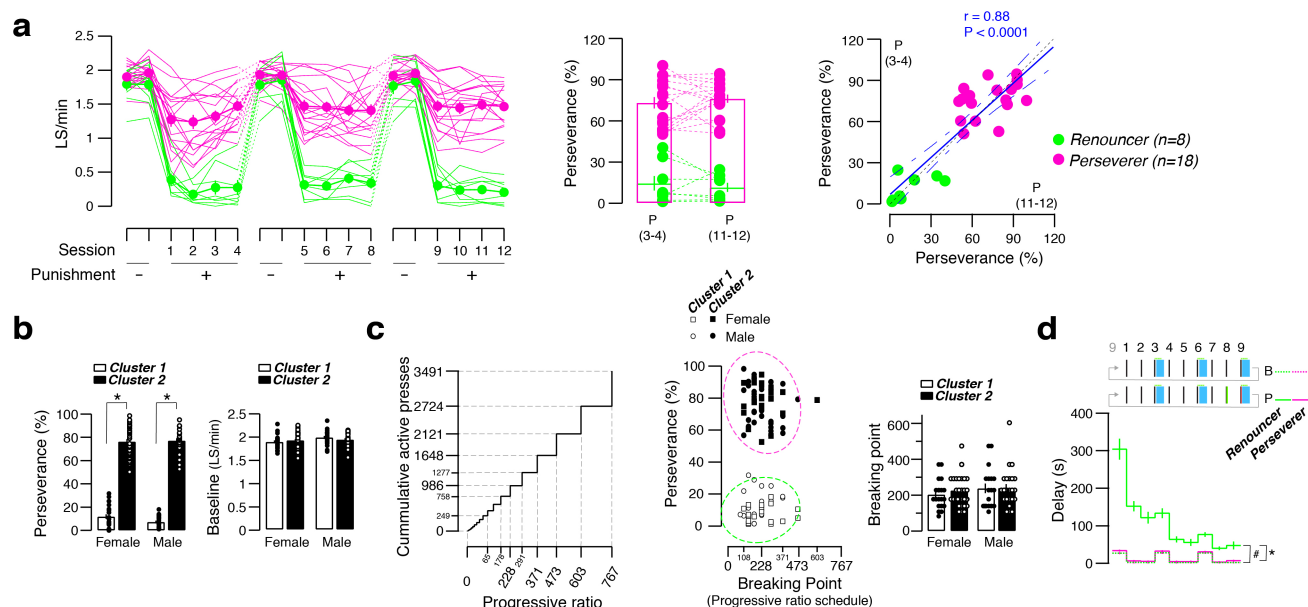
Extended Data Fig. 2 | Examples from two mice of lever presses and outcomes during oDASS. Raster plots for two mice, one keeping a high stimulation rate during punished sessions (example 2) and one falling

to a low stimulation rate during punished sessions (example 1). Every action and the associated outcome as a function of time is shown for three baseline (B) and four punished (P) sessions.



Extended Data Fig. 3 | Emergence of two clusters of mice with punished oDASS. a, Hierarchical clustering of the entire dataset (time-event for eight parameters during three baseline and four punished sessions). Each column corresponds to a mouse. Heat maps for delays between the active lever presses, the number of laser stimulation events, the time of the last laser stimulation, the time remaining in a trial and the number of inactive lever press in one mouse are plotted. Two clusters are found in the resulting dendrogram (green, renouncer; pink, perseverer). Vertical

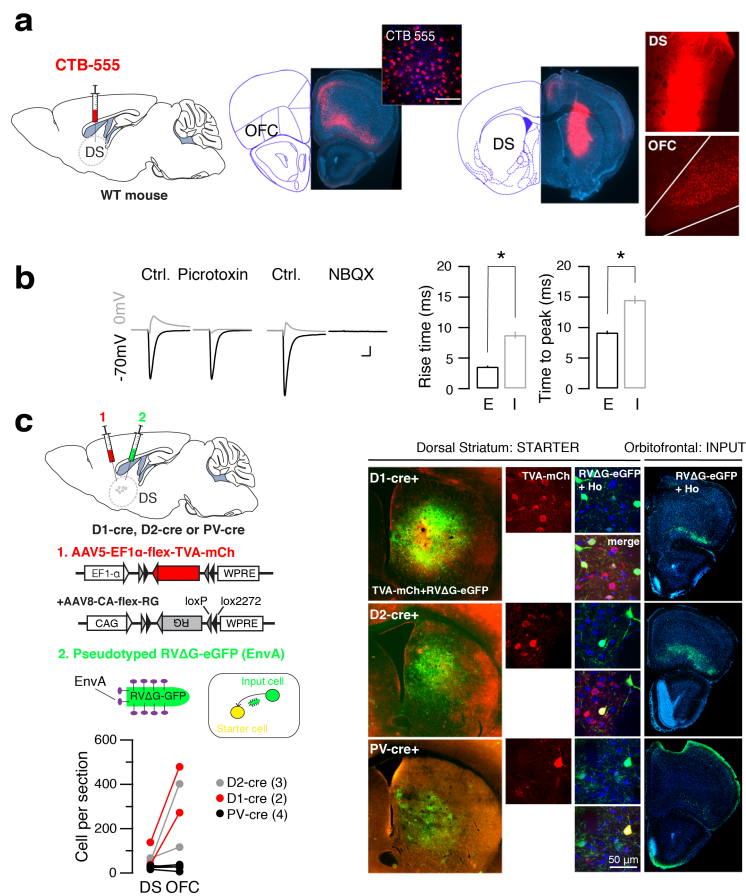
dashed lines separate renouncing from persevering mice. Heat map (greyscale) represents the perseverance (measured as the oDASS rate during punished sessions 3 and 4, normalized to the baseline sessions) of each mouse as a function of the clustering. **b,** Before clustering, we applied a nonlinear dimension reduction to project the high-dimensional dataset into a two-dimensional representation (left). Mapping the perseverance onto this map (right) shows that this variable can be used to categorize the mice as renouncing and persevering mice.



Extended Data Fig. 4 | Stability of perseverance and oDASS acquisition parameters.

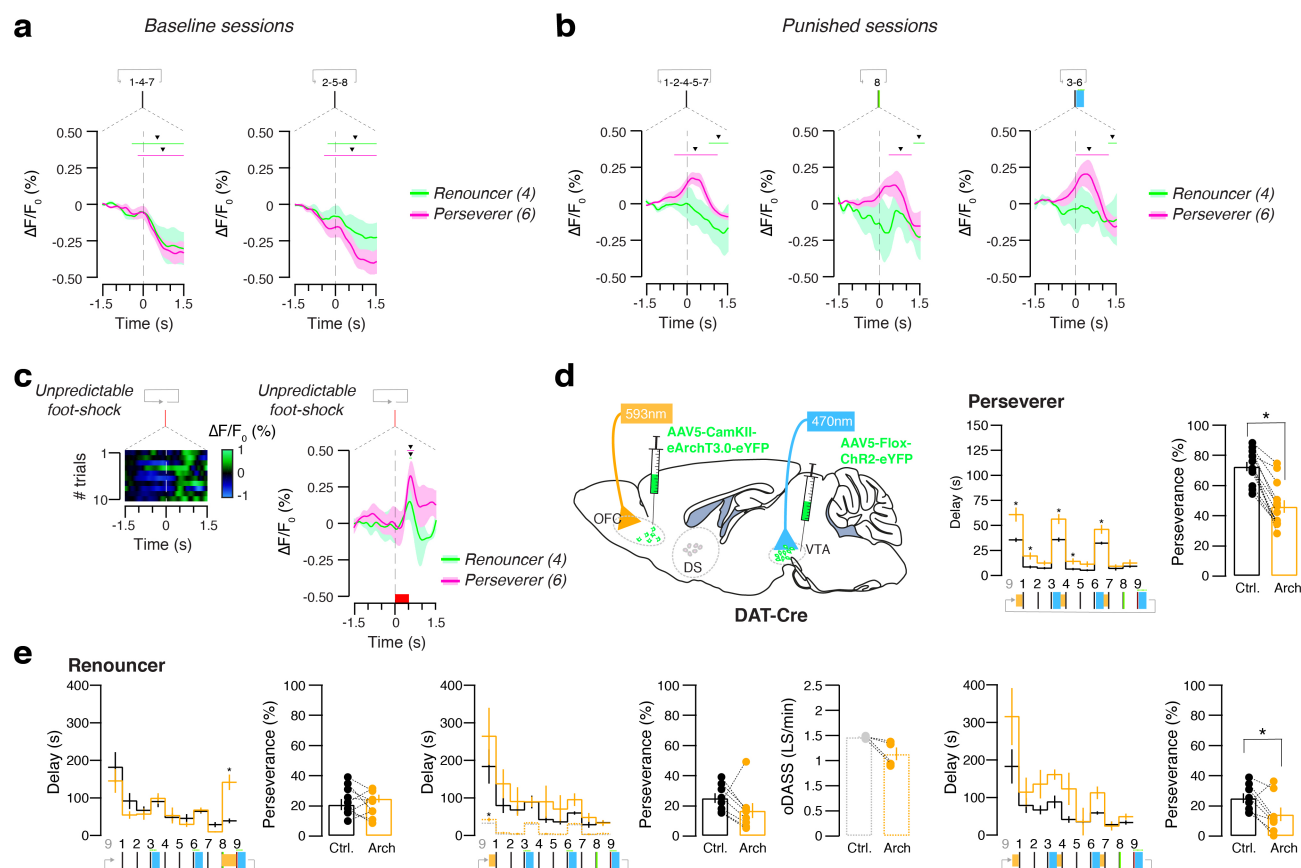
a, Three blocks of two baseline and four punished oDASS sessions were performed during a two-month period. Left, oDASS rate as a function of advancing sessions and average for the two groups. Middle, perseverance measured during punishment sessions 3–4 was compared to sessions 11–12. Perseverance was calculated as the average oDASS during the two punished sessions normalized to the corresponding baseline rate. No effect of punishment block was detected. Right, correlation between perseverance between first and last block (sessions 3–4 versus session 11–12), Pearson's $r = 0.88$ ($n = 26$). **b**, Left, perseverance and baseline rate for male and female mice of the two clusters (for perseverance, ANOVA followed two-sided t -test: $*P < 0.0001$, $t_{60} = 23.02$ for persevering versus renouncing female mice ($n = 39$ and 23 mice, respectively); $*P < 0.0001$, $t_{45} = 22.30$ for persevering versus renouncing male mice ($n = 20$ and 27 mice, respectively). Right, no difference in baseline rate was detected between the two clusters, nor between male and female mice

($n = 109$ mice). **c**, Left, cumulative active presses as a function of the progressive ratio. Middle, perseverance as a function of the breakpoint during the progressive ratio schedule for male (42, squares) and female (56, circles). Note that in Fig. 1e, data are presented with cumulative active presses, which avoids the steps that are observed inherent to the breakpoint plot. No difference in breakpoint was detected between the two clusters, nor between male and female mice ($n = 98$ mice). **d**, Temporal structure of an oDASS trial showing the delays between active press during baseline and punished sessions for cluster 1 and cluster 2 mice (renouncing and persevering mice, respectively). Delays were increased in renouncing mice (ANOVA followed by two-sided t -test: $*P < 0.05$ in punished sessions for persevering versus renouncing mice for every delay ($n = 66$ and 43 mice, respectively); $^{\#}P < 0.05$; baseline/punished for renouncing mice for every delay). Data are mean \pm s.e.m. See Supplementary Table 1 for complete statistics.



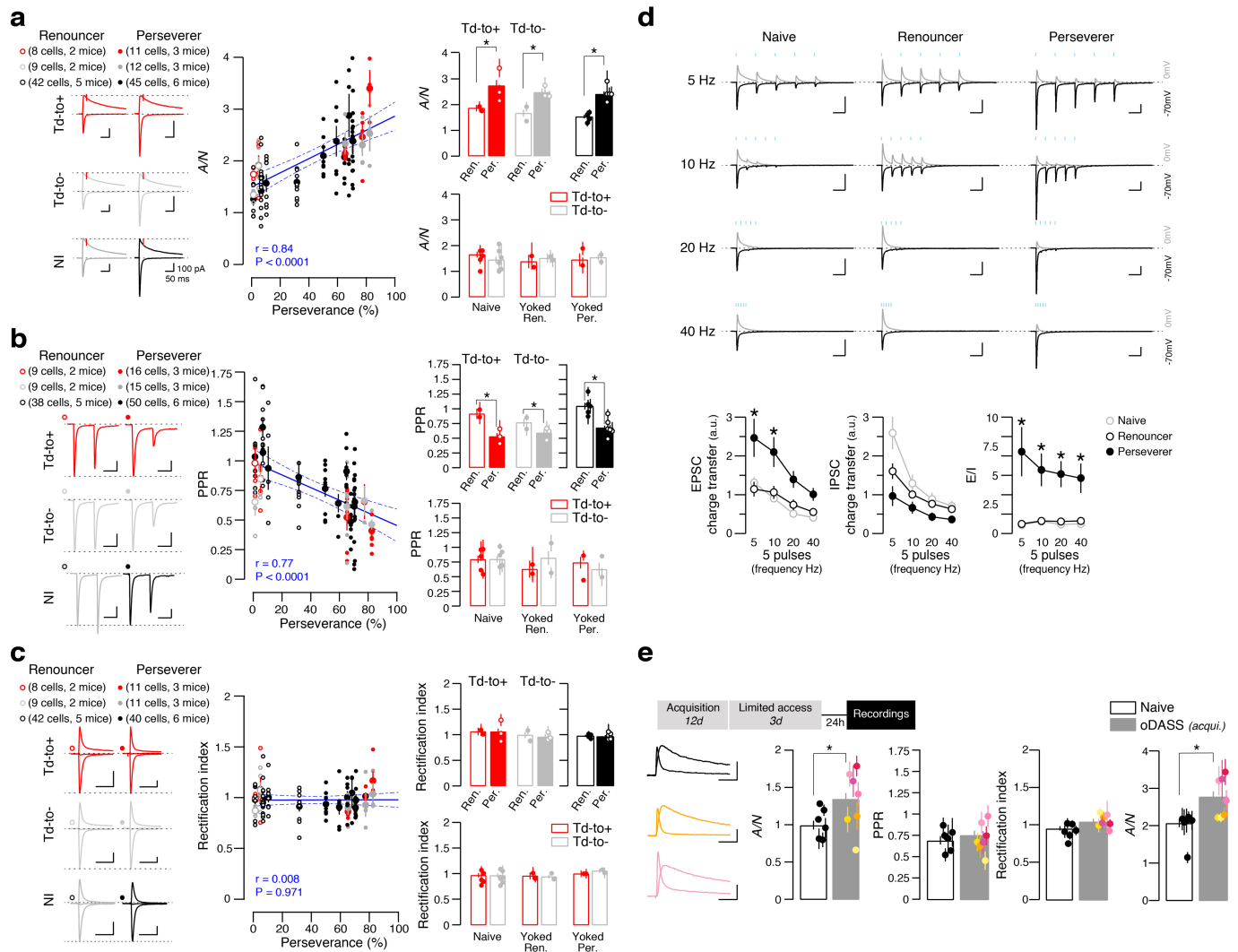
Extended Data Fig. 5 | Characterization of the connection between OFC and striatum. **a**, Retrograde tracing from the striatum with cholera toxin subunit B coupled to a red dye (CTB-555). Scale bar, 50 μ m. Experiment was repeated in $n = 8$ mice. **b**, OFC–striatum optogenetic stimulation and recordings of EPSCs, blocked by the AMPAR antagonist (NBQX) and IPSCs blocked by GABA $_A$ antagonist (picrotoxin). Rise time and time to peak for IPSCs and EPSCs indicate a feed-forward circuit between principal neurons of the OFC and striatum (two-sided paired t -test for IPSCs versus EPSCs: $t_{25} = 9.55$, $*P < 0.0001$ and $t_{25} = 7.57$, $P < 0.0001$ for rise time and time to peak, respectively, $n = 26$ cells from five mice). Scale bars, 20 ms, 100 pA. **c**, Schematic of retrograde tracing from specific cell types of the striatum using a rabies virus that was

injected in transgenic Cre-mouse lines encoding the dopamine D1 or D2 receptor and parvalbumin. A first injection (red) leads to cell-type-specific expression of the EnvA receptor TVA and the RG protein. After two weeks an EnvA-pseudotyped and glycoprotein (Δ G)-deleted rabies virus (EnvA and RV Δ G-GFP) is injected (green) and taken up by the cells that express TVA and thus turn yellow (starters). Trans-complemented with glycoprotein by infection of the AAV8-CA-flex-RG and RV Δ G-GFP transsynaptically caused a spread to upstream neurons (inputs). The injection site in the striatum (left) and high-magnification images show starter cells. Retrogradely infected neurons in the OFC at low and higher magnification (right). See Supplementary Table 1 for complete statistics. **a**, **c**, Images reproduced from Paxinos and Franklin⁴⁴, copyright © 2001.



Extended Data Fig. 6 | Activity and manipulation of OFC–striatum projection during oDASS. **a**, Calcium signal ($\Delta F/F_0$, mean \pm s.e.m.) around active press 1 and 2 of the FR3 schedule (numbers 1, 4, 7 and 2, 5, 8) during a baseline session for renouncing and persevering mice (green diamonds and pink bars indicate a significant deviation from baseline, $n = 4$ and 6 mice). **b**, Averaged calcium signal ($\Delta F/F_0$, mean \pm s.e.m.) around active press 1 and 2 of the FR3 schedule (all but press number 8), around the active press number 8 (leading to the shock-associated cue) and around the lever press that terminates the non-shock FRs in punished sessions for renouncing and persevering mice (green diamonds and pink bars indicate a significant deviation from baseline, $n = 4$ and 6 mice). **c**, Example of trial activity map of the calcium signal ($\Delta F/F_0$, mean \pm s.e.m.) around an unpredictable foot shock (500 ms, 0.25 mA, repeated in 10 times in one mouse). For each animal, 10 unpredictable foot shocks were delivered during a separate recording. Grouped data for the calcium signal ($\Delta F/F_0$, mean \pm s.e.m.) around an unpredictable foot shock for renouncing and persevering mice (green diamonds and pink bars indicate a significant deviation from baseline, $n = 4$ and 6 mice). See Supplementary Table 1 for statistics. **d**, Scheme of a mouse brain infected with eArchT3.0–eYFP in the OFC and with ChR2–eYFP in the VTA (left). For persevering mice, OFC inhibition with eArchT3.0

between oDASS and the next FR initiation (or for a maximum of 90 s) delayed the next press of persevering mice (ANOVA followed by two-sided t -test: $*P < 0.05$ when comparing control versus eArchT3.0 delays during punished sessions, $n = 13$ mice). Perseverance changed (from 73% to 46%) as a consequence of eArchT3.0 stimulation (two-sided paired t -test: $t_{12} = 9.13$, $*P < 0.0001$, $n = 13$ mice for control versus eArchT3.0 before each FR initiation). **e**, For renouncing mice, OFC inhibition with eArchT3.0, after the punishment-predictive cue, between punished oDASS and the next FR initiation (or for a maximum of 90 s) or between each oDASS and the next FR initiation slightly delayed the next press (ANOVA followed two-sided paired t -test: $*P < 0.05$ when comparing control and eArchT3.0 delays during punished sessions). Perseverance was reduced as a consequence of eArchT3.0 stimulation between each oDASS and the next FR initiation (two-sided paired t -test: $t_7 = 2.62$, $*P = 0.034$, $n = 8$ mice for control versus eArchT3.0 before each FR initiation). The oDASS rate during a baseline session was not significantly changed between punished oDASS and the next FR initiation (or for a maximum of 90 s) by inhibition with eArchT3.0. Data are mean \pm s.e.m. See Supplementary Table 1 for complete statistics. **d**, Line drawing modified from Paxinos and Franklin⁴⁴, copyright © 2007.

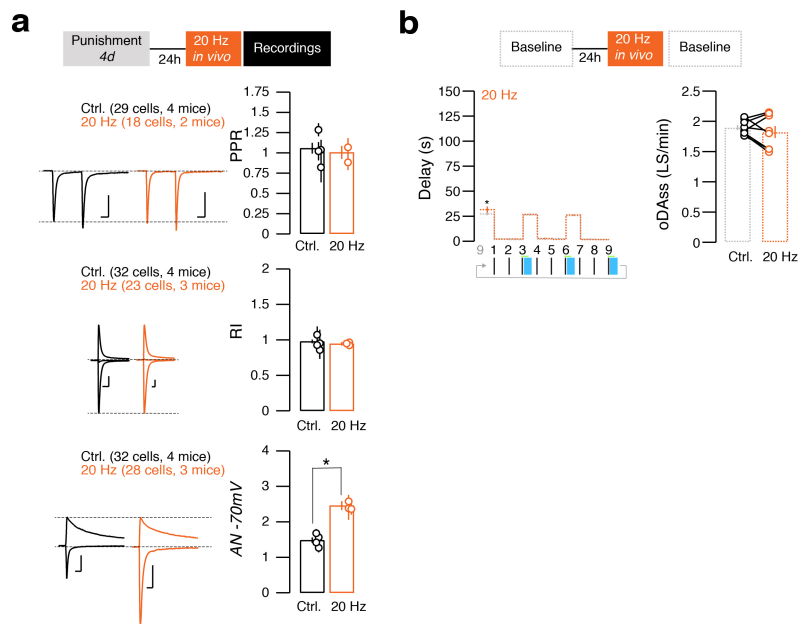


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Plasticity at OFC–striatum synapses correlates with compulsive oDASS.

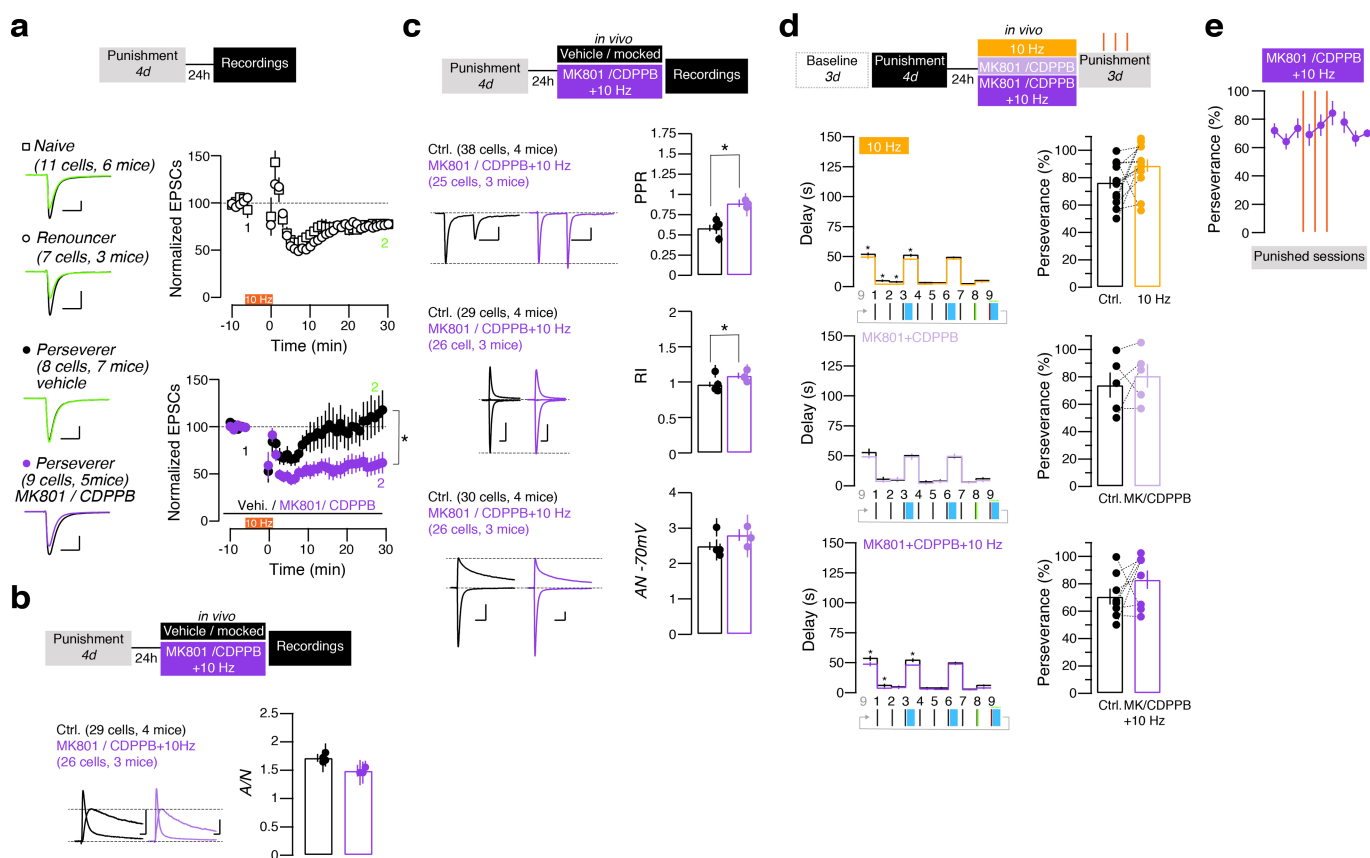
a, Average of 10 sweeps for AMPAR EPSCs recorded at -70 mV and for EPSCs recorded at $+40$ mV in slices from renouncing and persevering mice, with or without SPN identification with tdTomato. NMDA amplitude was analysed 20 ms after the peak of the EPSC recorded at $+40$ mV. The AMPAR/NMDAR ratio for every recorded neuron as a function of perseverance, data are mean \pm s.e.m. per animal and correlation (Pearson's $r = 0.84$, $P < 0.0001$; $n = 127$ cells from 16 mice). Data are mean AMPAR/NMDAR ratio for renouncing and persevering mice (ANOVA followed by two-sided t -test: $t_{17} = 3.13$, $*P = 0.007$ and $t_{19} = 3.53$; $*P = 0.002$ for tdTomato⁺ and tdTomato⁻, respectively ($n = 8$ and 11 mice; 9 and 12 cells); $t_{85} = 6.68$, $*P < 0.0001$, renouncing versus persevering for non-identified neurons (42 and 45 cells, respectively)). Data are mean AMPAR/NMDAR ratio for naive mice and for mice yoked to renouncing or persevering mice, with Drd1a–tdTomato identification (naive mice, $n = 14$ tdTomato⁺ cells from 5 mice and $n = 12$ tdTomato⁻ cells from 7 mice; mice yoked to renouncing mice, $n = 6$ tdTomato⁺ cells from 2 mice and $n = 9$ tdTomato⁻ cells from 2 mice; mice yoked to persevering mice, $n = 6$ tdTomato⁺ cells from 2 mice and $n = 8$ tdTomato⁻ cells from 2 mice). **b**, Average of 10 sweeps for AMPAR EPSCs recorded at -70 mV, with two short light pulses spaced with a 76-ms interval in slices from renouncing and persevering mice, with or without SPN identification using tdTomato. PPR is the ratio of the amplitudes of the second EPSC over the first. PPR for every recorded neuron as a function of perseverance, mean \pm s.e.m. per animal and correlation (Pearson's $r = -0.77$, $P < 0.0001$; $n = 137$ cells from 16 mice). Data are mean PPR for renouncing and persevering mice (ANOVA followed by two-sided t -test: $t_{23} = 3.97$, $*P = 0.0005$ and $t_{22} = 1.72$, $P = 0.183$ for tdTomato⁺ and tdTomato⁻, respectively ($n = 9$ and 15 mice; 9 and 16 cells); $t_{86} = 5.66$, $*P < 0.0001$, renouncing versus persevering for non-identified neurons ($n = 38$ and 50 cells, respectively)). Data are mean PPR for naive mice and for mice yoked to renouncing or persevering mice, with Drd1a–tdTomato identification (naive mice, $n = 13$ tdTomato⁺ cells from 5 mice and $n = 11$ tdTomato⁻ cells from 7 mice; mice yoked to renouncing mice, $n = 6$ tdTomato⁺ cells from 2 mice and $n = 10$ tdTomato⁻ cells from 2 mice; mice yoked to persevering mice, $n = 7$ tdTomato⁺ cells from 2 mice and $n = 8$ tdTomato⁻ cells from 2 mice). **c**, Left, average of 10 sweeps for AMPAR EPSCs recorded at -70 , 0 and $+40$ mV in slices from renouncing and persevering mice, with or without SPN identification using tdTomato. The rectification of the AMPAR EPSCs was calculated as the ratio of the chord conductance calculated at negative potential divided by chord conductance at positive

potential. Middle, rectification index for every recorded neuron as a function of perseverance, mean \pm s.e.m. per animal and correlation (Pearson's $r = 0.008$, $P = 0.971$; $n = 121$ cells from 16 mice). Right, rectification index for renouncing and persevering in tdTomato⁺ or tdTomato⁻ cells (8 and 11 mice; 9 and 11 cells) and in unidentified SPNs (42 and 40 cells, respectively). Data are mean rectification index for naive mice and for mice yoked to renouncing or persevering mice with Drd1a–tdTomato identification (naive mice, $n = 13$ tdTomato⁺ cells from 5 mice and $n = 11$ tdTomato⁻ cells from 7 mice; mice yoked to renouncing mice, $n = 6$ tdTomato⁺ cells from 2 mice and $n = 9$ tdTomato⁻ cells from 2 mice; mice yoked to persevering mice, $n = 7$ tdTomato⁺ cells from 2 mice and $n = 8$ tdTomato⁻ cells from 2 mice). **d**, Average of 10 sweeps for EPSCs recorded at -70 mV and IPSCs recorded at 0 mV in slices from renouncing and persevering mice and in slices from naive mice. Five pulses were given at different frequencies (5 , 10 , 20 and 40 Hz) and the charge transfer was measured (area under the curve). The excitatory/inhibitory ratio (E/I) was calculated as the ratio charge transfer for EPSCs over IPSCs. The charge transfer of EPSCs was higher in slices from persevering mice at low frequencies (ANOVA followed by two-sided t -test: $t_{27} = 4.75$, $*P < 0.0001$; $t_{27} = 3.75$, $*P = 0.0007$; $t_{27} = 2.37$, $P = 0.057$; $t_{27} = 1.64$, $P = 0.306$ for 5 , 10 , 20 and 40 Hz, respectively (17 and 12 cells)). The charge transfer of IPSCs was not different between persevering and renouncing mice (17 and 12 cells, respectively). The ratio of charge transfer for EPSCs over IPSCs was higher in slices from persevering mice (ANOVA followed by two-sided t -test: $t_{27} = 6.22$, $*P < 0.0001$; $t_{27} = 4.39$, $*P < 0.0001$; $t_{27} = 4.07$, $P = 0.0002$; $t_{27} = 3.67$, $P = 0.001$ for 5 , 10 , 20 and 40 Hz, respectively (17 and 12 cells)). Measurements were obtained from four renouncing mice, three persevering mice and three naive mice. **e**, Average of 10 sweeps for AMPAR EPSCs in the presence of D-AP5 (50 μ M) and NMDAR EPSCs isolated by subtraction for oDASS mice and for naive mice. Mean AMPAR/NMDAR ratio, PPR and rectification index for naive and oDASS mice ($*P < 0.05$ for t -test comparing naive/oDASS). Each dot represents the mean \pm s.e.m. for all cells obtained in a given mouse. Recordings were obtained from six naive mice and seven oDASS mice (PPR: 48 cells from oDASS mice compared to 29 cells from naive mice; rectification index: 52 cells from oDASS mice compared to 24 cells from naive mice; AMPAR/NMDAR ratio with pharmacological isolation: 42 cells from oDASS mice compared to 23 cells from naive mice; AMPAR/NMDAR ratio without pharmacological isolation: 52 cells from oDASS mice compared to 26 cells from naive mice). Scale bars, 200 pA, 50 ms. Data are mean \pm s.e.m. See Supplementary Table 1 for complete statistics.



Extended Data Fig. 8 | Synaptic properties in persevering mice after 20-Hz stimulation in vivo. **a**, Average traces for EPSCs recorded to determine PPR, rectification index (RI) and AMPAR/NMDAR ratio without pharmacological isolation. Ex vivo measurement of the PPR, rectification index and AMPAR/NMDAR ratio after in vivo stimulation of OFC–striatum terminals at 20 Hz for 1 min in renouncing mice (for PPR: control versus 20 Hz, two-sided t -test: $t_{45} = 0.48$, $P = 0.63$, $n = 29$ and 18 cells, respectively; for rectification index: control versus 20 Hz: $t_{53} = 0.90$, $P = 0.37$, $n = 32$ and 23 cells respectively; for AMPAR/

NMDAR ratio: control versus 20 Hz: $t_{58} = 6.79$, $*P < 0.0001$, $n = 32$ and 28 cells, respectively). **b**, Effect of 20-Hz stimulation of OFC–striatum prior to a baseline session ($n = 8$ mice). Delay to engage the next action was not changed (ANOVA followed by two-sided t -test: $*P < 0.05$ when comparing control and 20 Hz). oDASS rate was not modified by 20 Hz stimulation before a baseline session ($t_7 = 0.97$, $P = 0.36$, $n = 8$ mice for control versus 20 Hz before a baseline session). Data are mean \pm s.e.m. See Supplementary Table 1 for complete statistics.

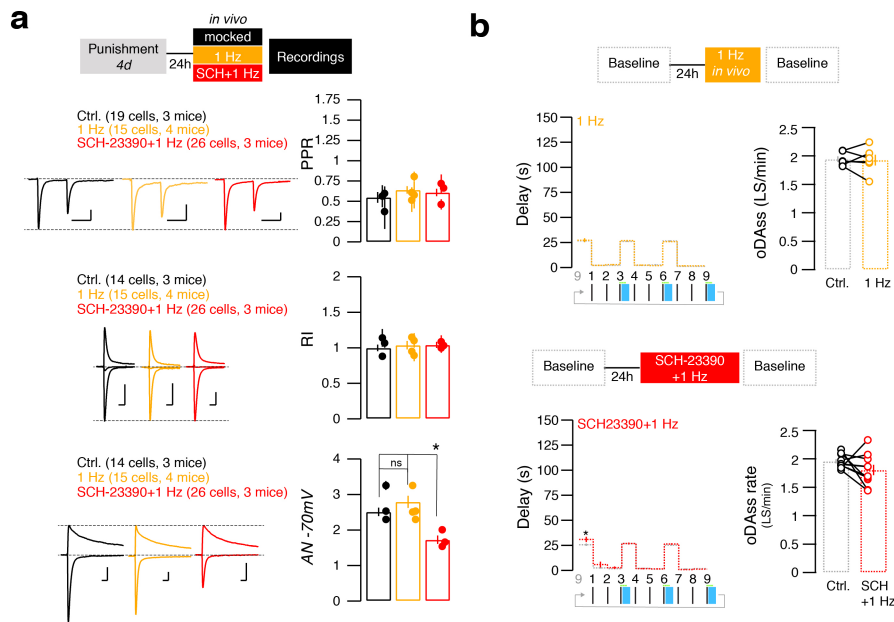


Extended Data Fig. 9 | Absence of an effect on compulsive behaviour by normalization of release probability in persevering mice.

a, Average traces for EPSCs recorded immediately before and 30 min after the LTD protocol (10 Hz for 5 min) and grouped data. In slices from persevering mice, LTD is unmasked by bath application of mGluR5 PAM (CDPPB, 100 μ M) and MK801 (NMDAR blocker, 10 μ M) (two-sided t -test: $t_{15} = 2.89$, $*P = 0.037$, $n = 8$ and 9 cells, respectively).

b, AMPAR/NMDAR ratio was left unchanged by in vivo stimulation of OFC–striatum terminals with 10 Hz in presence of MK801 and CDPPB (0.3 and 30 mg kg^{-1} , respectively) in persevering mice (two-sided t -test: $t_{53} = 1.86$, $P = 0.069$ for control versus 10 Hz treated with MK801 and CDPPB ($n = 29$ and 26 cells, respectively)). **c**, PPR was normalized by in vivo stimulation of OFC–striatum terminals with 10 Hz in the presence of MK801 and CDPPB in persevering mice (two-sided t -test: $t_{61} = 4.94$, $P < 0.0001$ for control versus 10 Hz with MK801 and CDPPB ($n = 38$ and 25 cells, respectively)). The rectification index was different between controls and mice treated in vivo with the stimulation of OFC–striatum terminals at 10 Hz in the presence of MK801 and CDPPB (two-sided t -test:

$t_{53} = 2.25$, $*P = 0.029$ for control versus 10 Hz with MK801 and CDPPB ($n = 29$ and 26 cells, respectively)). AMPAR/NMDAR ratio without pharmacological isolation was not different between controls and mice treated in vivo with the stimulation of OFC–striatum terminals at 10 Hz in the presence of MK801 and CDPPB (two-sided t -test: $t_{54} = 1.57$, $P = 0.12$ for control versus 10 Hz with MK801 and CDPPB ($n = 30$ and 26 cells, respectively)). **d**, Plots for delay between lever presses in punished sessions 12 h after 10 Hz, MK801 and CDPPB or 10 Hz with MK801 and CDPPB (ANOVA followed by two-sided paired t -test: $*P < 0.05$ when comparing control/treatment delays during punished sessions). Perseverance is not modified by any treatment (two-sided t -test: $t_{12} = 2.12$, $P = 0.056$, $n = 13$ mice for control versus 10 Hz; $t_4 = 0.73$, $P = 0.51$, $n = 5$ mice for control versus MK801 and CDPPB; $t_7 = 1.31$, $P = 0.231$, $n = 8$ mice for control versus 10 Hz with MK801 and CDPPB). **e**, During additional punished sessions without renewal of the intervention, perseverance remained unchanged ($n = 8$ mice). Data are mean \pm s.e.m. See Supplementary Table 1 for complete statistics.



Extended Data Fig. 10 | Effects of SCH23390 and 1 Hz in vivo in persevering mice. **a**, Average traces of EPSCs recorded to determine PPR, rectification index and AMPAR/NMDAR ratio without pharmacological isolation. Ex vivo measurement of the PPR, rectification index and AMPAR/NMDAR ratio after in vivo stimulation of OFC–striatum terminals at 1 Hz for 5 min in persevering mice, in the presence or absence of SCH23390 (ANOVA followed by two-sided *t*-test: for PPR: $t_{32} = 1.10$, $P = 0.83$; $t_{39} = 0.64$, $P > 0.99$; and $t_{43} = 0.57$, $P > 0.99$ for control versus 1 Hz ($n = 19$ and 15 cells, respectively); 1 Hz versus 1 Hz with SCH23390 ($n = 15$ and 26 cells, respectively); for rectification index: $t_{27} = 0.74$, $P > 0.99$; $t_{39} = 0.17$, $P > 0.99$; and $t_{38} = 1.00$, $P = 0.97$ for control versus 1 Hz ($n = 14$ and 15 cells, respectively); 1 Hz versus 1 Hz with SCH23390 ($n = 15$ and 26 cells, respectively); and control versus 1 Hz with SCH23390

($n = 14$ and 26 cells, respectively); for AMPAR/NMDAR ratio: $t_{27} = 1.64$, $P = 0.32$; $t_{39} = 5.93$, $P < 0.0001$; and $t_{38} = 3.96$, $P = 0.0007$ for control versus 1 Hz ($n = 14$ and 15 cells, respectively); 1 Hz versus 1 Hz with SCH23390 ($n = 15$ and 26 cells, respectively); and control versus 1 Hz with SCH23390 ($n = 14$ and 26 cells, respectively)). **b**, Delay to engage the next action was not changed (ANOVA followed by two-sided paired *t*-test: $*P < 0.05$ when comparing control versus 1 Hz or control versus 1 Hz with SCH23390, $n = 10$ mice). oDASS rate was not modified by 1 Hz or 1 Hz with SCH23390 prior to a baseline session ($t_5 = 0.22$, $P = 0.84$, $n = 6$ mice for control versus 1 Hz before a baseline session and $t_9 = 1.48$, $P = 0.17$, $n = 10$ mice for control versus 1 Hz with SCH23390 before a baseline session). Data are mean \pm s.e.m. See Supplementary Table 1 for complete statistics.

Structure of native lens connexin 46/50 intercellular channels by cryo-EM

Janette B. Myers^{1,4}, Bassam G. Haddad^{1,4}, Susan E. O'Neill¹, Dror S. Chorev², Craig C. Yoshioka³, Carol V. Robinson², Daniel M. Zuckerman³ & Steve L. Reichow^{1*}

Gap junctions establish direct pathways for cell-to-cell communication through the assembly of twelve connexin subunits that form intercellular channels connecting neighbouring cells. Co-assembly of different connexin isoforms produces channels with unique properties and enables communication across cell types. Here we used single-particle cryo-electron microscopy to investigate the structural basis of connexin co-assembly in native lens gap junction channels composed of connexin 46 and connexin 50 (Cx46/50). We provide the first comparative analysis to connexin 26 (Cx26), which— together with computational studies— elucidates key energetic features governing gap junction permselectivity. Cx46/50 adopts an open-state conformation that is distinct from the Cx26 crystal structure, yet it appears to be stabilized by a conserved set of hydrophobic anchoring residues. ‘Hot spots’ of genetic mutations linked to hereditary cataract formation map to the core structural–functional elements identified in Cx46/50, suggesting explanations for many of the disease-causing effects.

Cell-to-cell communication directed by gap junctions is essential to neuronal function and cardiac coupling, and for coordinating intercellular signalling and metabolic activity in most tissues (for example, heart, skin, liver and eye lens)¹. Genetic mutation or aberrant regulation of gap junctions is linked to a variety of pathological conditions, including cardiac arrhythmia, stroke, blindness, deafness, skin disease and cancers^{2–4}.

Intercellular channel formation occurs through an assembly of twelve connexin subunits⁵. Within the plasma membrane, six connexins are organized into a hemichannel structure. Hemichannels from neighbouring cells dock together to form complete cell-to-cell channels, which cluster to form large gap junction plaques. A remarkably large channel pore provides passage to diverse chemical messages; these include ions, metabolites, hormones and other small signalling molecules less than about 1 kDa in size (for example, K⁺, cyclic AMP (cAMP), inositol triphosphate (Ins(1,4,5)P₃) and glucose). In this way, interconnected cells can exchange electrical and chemical information across an entire tissue or organ.

Humans express 21 connexin isoforms in a cell-type-specific fashion⁶. Most cells express multiple isoforms, and certain connexins display an ability to co-assemble, either by docking two hemichannels composed of different isoforms (heterotypic) or through mixed isoform assembly within the same hemichannel (heteromeric). This complexity is thought to allow cells to fine-tune the conductance of chemical messages and support coupling across different cell types⁷. However, our understanding of the physical basis of connexin isoform compatibility, conductance, substrate selectivity and channel gating remains limited^{8,9}, as high-resolution structural information obtained by crystallographic analysis has so far been restricted to just a single model system, Cx26^{10,11}.

To gain further insight into the mechanistic effects of gap junction isoform diversity and heteromeric assembly, we applied single-particle imaging methods by cryo-electron microscopy (cryo-EM) to elucidate the structure of native channels made up of Cx46 and Cx50 (Cx46/50), isolated from the eye lens, in which connexin-mediated communication

is required for growth, differentiation and maintenance of lens transparency to support vision¹². Comparative molecular dynamics simulations reveal key features of ion permeation and selectivity, and suggest that Cx46/50 adopts a more stable open-state conformation compared to the previously described Cx26 crystal structure¹⁰.

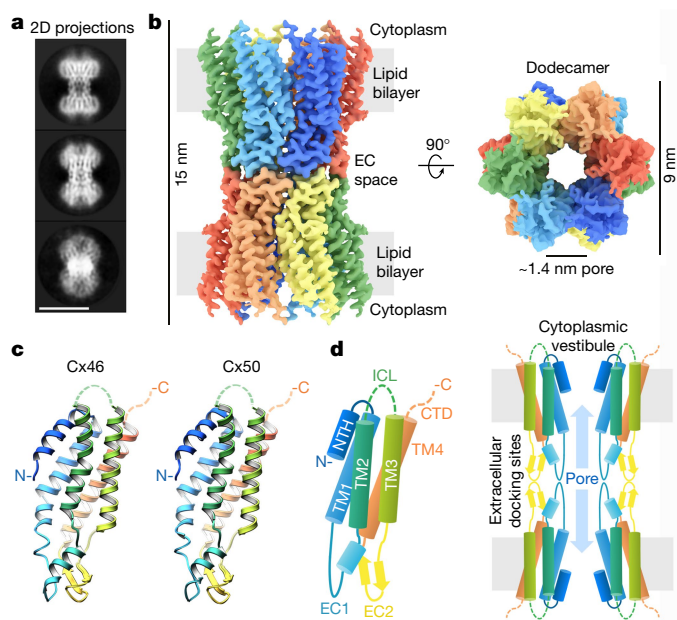
Structural overview of Cx46/50

Cx46 and Cx50 form intercellular channels in the mammalian lens, which are potentially heteromeric or heterotypic^{13,14}. We isolated native Cx46/50 intercellular channels from core lens tissue (sheep Cx44/49), and verified heteromeric co-assembly by biochemical analysis and chemical cross-linking mass spectrometry (Extended Data Fig. 1). The structure of these Cx46/50 intercellular channels was resolved by single-particle cryo-EM to near-atomic resolution (3.4 and 3.5 Å, from two independent datasets) (Fig. 1a, b, Extended Data Figs. 1–3). The resulting density maps revealed a 15-nm-long dodecameric (12-mer) channel with a girdled waist (about 6–9-nm wide). There is a large unobstructed pore, approximately 1.4 nm in diameter, along the channel axis, consistent with the proposed open-state conformation (Fig. 1b).

We were unable to resolve a specific pattern of Cx46/50 heteromeric or heterotypic co-assembly using 3D classification or refinement strategies (Methods, Extended Data Figs. 4, 5). Nevertheless, high-resolution features corresponding to side-chain densities are observed throughout the reconstructions following 12-fold symmetry refinement (that is, by averaging signal contributed by both Cx46 and Cx50). Therefore, these two isoforms, which share about 80% core-sequence identity (88% similarity), also share a highly similar 3D structure (Fig. 1b, c, Extended Data Figs. 4–6), consistent with the ability of Cx46/50 to co-assemble in a variety of heteromeric and/or heterotypic states.

Atomic models for Cx46 and Cx50 were built into the averaged cryo-EM density map, and various heteromeric and heterotypic channels were constructed for comparative analysis. Whereas the structures display excellent validation statistics (Extended Data Fig. 2), local resolution assessment of the atomic models and experimental

¹Department of Chemistry, Portland State University, Portland, OR, USA. ²Physical and Theoretical Chemistry Laboratory, University of Oxford, Oxford, UK. ³Department of Biomedical Engineering, Oregon Health and Science University, Portland, OR, USA. ⁴These authors contributed equally: Janette B. Myers, Bassam G. Haddad. *e-mail: reichow@pdx.edu



density identified features of both models that were less well-defined by the density map, in particular at sites at which the two isoforms differ in sequence (Extended Data Figs. 3–5, Supplementary Tables 2, 3). Analysis of the presented models should be approached with caution owing to intrinsic limitations of our heterogeneous dataset, which may extend beyond local differences in primary sequence; for example, owing to the possibility of one isoform being more well-ordered and potentially biasing interpretation.

The refined Cx46/50 structures comprise four alternating transmembrane α -helices (TM1–TM4), two extracellular domains (EC1 and EC2) connecting TM1–TM2 and TM3–TM4, respectively, and an N-terminal helix (NTH) domain that folds into the channel vestibule and is connected to the pore-lining TM1 helix via a short linker (Fig. 1c, d). Density for Met1 is not observed in the cryo-EM maps and was shown, by tandem mass spectrometry (MS/MS), to be removed in both Cx46 and Cx50. The resulting N-terminal glycine (G2) is partially acetylated (Extended Data Fig. 1), as shown for the bovine isoforms^{15,16}. The intracellular loop (ICL) connecting TM2–TM3 and the cytoplasmic C-terminal domain (CTD) containing the native cleavage sites of Cx46 and Cx50 are also not resolved¹⁶. The ICL and CTD were also not observed in the crystallographic structures of Cx26^{10,11}, presumably owing to intrinsic disorder in these regulatory domains.

The close structural similarity between Cx46 and Cx50 results in highly similar interfacial interactions that include conserved regions of hydrophobic packing over the transmembrane region, and a highly similar hydrogen-bond–ion-pair network between adjacent (homomeric interface; Fig. 2a, b, Extended Data Fig. 6) and opposed subunits (heterotypic interface; Fig. 2c, d, Extended Data Fig. 6). Most of these stabilizing interactions are present in Cx26^{10,17}, including the EC1 Q/N motif (Fig. 2c) and the EC2 pairing involving the K/R–N–D motif (Fig. 2d), a conserved element among group I heterotypic compatible isoforms¹⁸. Although Cx46/50 and Cx26 are not classified as

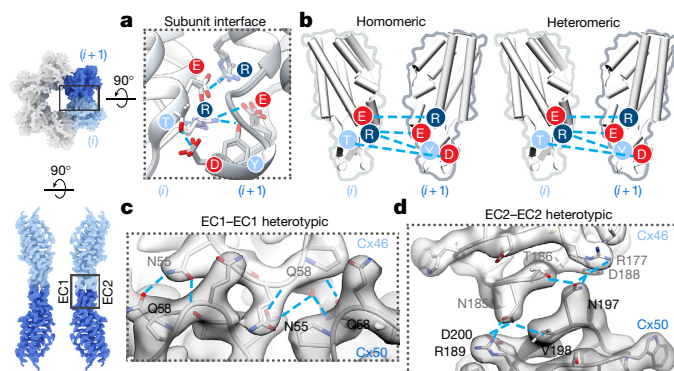


Fig. 2 | Heteromeric and heterotypic interactions between Cx46 and Cx50. **a**, Structural overlay (**a**) and illustration (**b**) of the conserved electrostatic network between neighbouring subunits (i and $i + 1$) identified in homomeric Cx46 and Cx50 and heteromeric Cx46/50 models. Labels are coloured according to amino acid charge characteristics (red, negative; dark blue, positive; light blue, polar). **c**, **d**, Magnified view of EC1–EC1 (**c**) and EC2–EC2 (**d**) docking-site interactions, with atomic models of Cx46 and Cx50 hemichannels built into the cryo-EM density in a heterotypic configuration. Conserved amino acids involved in hydrogen-bond pairing (cyan lines) are labelled.

heteromeric compatible channels¹⁹, the conserved features at the heteromeric interface are congruent with the current understanding that heteromeric co-assembly of connexins is established during biogenesis in the ER–Golgi network¹⁹.

Overall, despite significant sequence differences, Cx46 and Cx50 (α -family connexins) display core structural features that are very similar to the β -family member, Cx26¹⁰ (pairwise C_{α} root mean squared deviation (r.m.s.d.) = 2.18 Å and 2.14 Å versus Cx46 and Cx50, respectively). These different connexin family members thus share a conserved connexin fold and gap junction channel architecture, as presented in Fig. 1d, and our structures are consistent with early low-resolution electron diffraction studies on Cx43 obtained in a lipid bilayer^{20,21}. Despite these general similarities, however, we uncovered substantial differences between Cx46/50 and Cx26 localized to key functional sites, which we expect to contribute to isoform-specific permeation and selectivity properties and provide insight into the interactions responsible for fully stabilizing the open-state conformation of these channels, detailed below.

Energetics of ion permeation and selectivity

Comparisons between Cx50, Cx46 and Cx26 intercellular channels reveal distinct electrostatic pore pathways, with shared regions of negative charge potential and steric constriction sites formed by the NTH domains that narrow the cytoplasmic vestibule to around 10–12 Å at both ends of the channel (Fig. 3a, b). The pore diameters are within the range determined for other connexin channels²², and fitting with the general ability of gap junctions to enable a variety of molecules of less than about 1 kDa in size (such as those in Fig. 3b) to cross between cells. However, these channels can display a substantial level of isoform-specific selectivity for molecules below this size cut-off, including discrimination between small charged ions²³.

To validate our structural models and gain insight into the mechanism of ion selectivity, we conducted comparative all-atom molecular dynamics simulations and potential-of-mean-force (PMF) calculations to define the free-energy landscape of potassium (K^{+}) and chloride (Cl^{-}) permeation for Cx50, Cx46 and Cx26 (Fig. 3c, d, Extended Data Figs. 7, 8). PMFs obtained for Cx26 should be interpreted cautiously owing to significant dynamical behaviour observed for the NTH domain during molecular dynamics simulations (Extended Data Fig. 7), described in detail in the following section.

Cx50, Cx46 and Cx26 form high-conductance ion channels, with preference for conductance of cations over anions. For molecular dynamics simulation, the N terminus of each of the models was

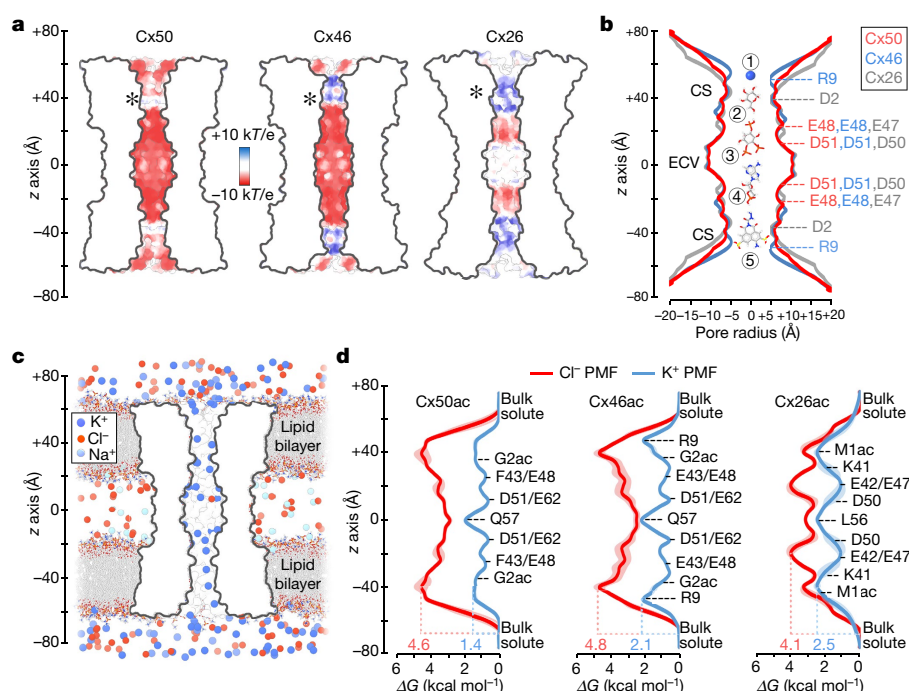


Fig. 3 | Comparative pore profile and energetics of ion permeation. **a**, Cut-away surface representation of Cx50 (left), Cx46 (centre) and Cx26 (right) (PDB 2ZW3¹⁰; residues 2–96 and 132–217), coloured by coulombic potential (red, negative; white, neutral; blue, positive). *T*, temperature; *k*, Boltzmann constant; *e*, charge of an electron; *, location of N terminus. **b**, Pore radius determined for experimental structures of Cx50 (red), Cx46 (blue) and Cx26 (grey, PDB 2ZW3¹⁰). Locations of constriction sites (CS) and ECV are indicated. Structures of representative substrates are displayed to scale: K⁺ ion (1), glucose (2), Ins(1,4,5)P₃ (3), cAMP (4) and Lucifer yellow (5). **c**, Snapshot of the Cx50 molecular dynamics simulation,

showing the membrane channel (white) embedded in two lipid bilayers and solvated in the presence of intracellular K⁺, and extracellular Na⁺ and Cl[−] ions. Water molecules not shown. **d**, PMF describing the free-energy landscape (ΔG) experienced by K⁺ ions (blue trace) and Cl[−] ions (red trace) permeating the channel pore. Symmetrized values are shown for acetylated models of Cx50ac (left), Cx46ac (centre) and Cx26ac (right), with non-symmetrized values in lighter shading. Amino acid positions are presented for correlation purposes, and do not represent deconvolution of the free-energy components.

acetylated (Cx50ac, Cx46ac and Cx26ac), as this form is expected to represent the predominant species *in vivo*²⁴, and previous molecular dynamics studies suggest this co-translational modification is required to obtain physiologically relevant charge selectivity of Cx26²². PMFs for K⁺—the major permeant ion—reveal peak energetic barriers within the constriction site, ranging from 1.4 kcal mol^{−1} for Cx50ac to 2.1 kcal mol^{−1} in Cx46ac and 2.5 kcal mol^{−1} in Cx26ac (Fig. 3d). These relatively low barriers are similar to the peak energetic barrier determined for other high-conductance Na- and K-channels (~2–3 kcal mol^{−1})^{25–27}, and are consistent with the range of experimental unitary conductance values of these channels (around 220 pS for Cx50²⁸ versus 140–135 pS for Cx46²⁹ and Cx26³⁰, in 130–140 mM CsCl).

The differences in K⁺ PMF correlate with isoform-specific differences in both steric and electrostatic environments. The constriction site of Cx50ac displays the lowest barrier and is characterized by a nearly completely electronegative coulombic potential, owing in part to neutralization of the N terminus by acetylation (shown by the asterisk in Fig. 3a, Extended Data Fig. 8). The major K⁺ energy barrier of Cx46ac correlates with the position of the positively charged residue R9 (*z* ≈ 50 Å, in which *z* is distance along the pore axis) (Fig. 3d, Extended Data Fig. 8), which also limits the constriction site of Cx46 to about 10 Å in our model (versus about 12 Å for Cx50) (Fig. 3b). However, the cryo-EM density map is not well-defined at this site (Supplementary Table 2), probably because (at least in part) of the conformational flexibility of this residue, as dynamical behaviour is observed during molecular dynamics simulation. These dynamics of R9 effectively modulate the steric barrier of Cx46 (between about 10 and 12 Å). The constriction-site K⁺ energy barrier of Cx26ac correlates with the location of the basic residue K41 (*z* ≈ 50 Å) (Fig. 3d), located on TM1 just below the NTH domain, as previously reported^{22,31}.

Free-energy minima for K⁺ are localized within the extracellular vestibule (ECV; *z* ≈ 10 Å and *z* ≈ 30 Å) of all three isoforms (Fig. 3d), supporting the role of EC1 in establishing charge selectivity and conductance^{31–33}. In Cx46 and Cx50, several negatively charged residues (for example, E48, D51 and E62) localize with regions of high K⁺ ion density (Fig. 3c, d, Extended Data Fig. 9). Notably, charge substitutions at D51 resulted in decreased unitary conductance in Cx46 hemichannels³⁴. E48 and D51 are conserved in Cx26 (equivalent to E47 and D50) (Fig. 3d), and establish transient binding interactions with K⁺ ions during molecular dynamics simulation. These sites have also been implicated in Ca²⁺ regulation in Cx26 by X-ray crystallography¹¹, molecular dynamics studies^{35,36} and by functional mutation studies of Cx46³⁶. Therefore, competitive K⁺ binding at these sites may contribute to the mechanism of Ca²⁺ regulation or sensitivity. E62 (in Cx46/50) appears to form an additional cation-binding site, through coordination between the carboxylate side chain and nearby backbone-carbonyl oxygens (Extended Data Fig. 9). E62 is not conserved in other human connexin isoforms (with the exception of Cx43), and may therefore constitute an isoform-specific regulatory site. Extracellular Ca²⁺ is involved in the mechanism of closing (or gating) connexin hemichannels³⁷, and competition by K⁺ binding at this putative site may contribute to the mechanism of potentiation of Cx50 and Cx46 hemichannels by extracellular K⁺ ions³⁸.

Cx50, Cx46 and Cx26 display an appreciable level of selectivity towards positively charged small ions, with permeability ratios of K⁺ to Cl[−] (P_{K^+}/P_{Cl^-}) ranging between around 2.5 to 10 (refs. ^{28,32,39–41}). Hydrated K⁺ and Cl[−] ions, with a diameter of about 7 Å, would pass unobstructed through a 10–12 Å steric constriction site; yet for all three isoforms, the peak energy barriers to Cl[−] are considerably larger than for K⁺ (Fig. 3d, Extended Data Fig. 8). Peak Cl[−] barriers localize within the constriction-site region of Cx46ac and Cx50ac

(4.8 kcal mol⁻¹, $z \approx 40$ Å; and 4.6 kcal mol⁻¹, $z \approx 50$ Å, respectively), and slightly deeper into the channel pore for Cx26ac, near the constriction-site–ECV border (4.1 kcal mol⁻¹, $z \approx 20$ Å). As a proxy for degree of P_{K^+}/P_{Cl^-} selectivity, we assessed the difference in peak K^+ and Cl^- PMF barriers ($\Delta\Delta G = 3.2$ kcal mol⁻¹ for Cx50ac, 2.7 kcal mol⁻¹ for Cx46ac and 1.6 kcal mol⁻¹ for Cx26ac). These relatively small differences in free energy are consistent with their moderate P_{K^+}/P_{Cl^-} selectivity ratios, and on the order of those defined for bacterial sodium channels (around 3.0–3.5 kcal mol⁻¹)^{26,27}, which display only modest selectivity for Na^+ over K^+ ($P_{Na^+}/P_{K^+} \approx 10$ –30). By contrast, voltage-gated K^+ channels display almost ideal selectivity for K^+ over Na^+ ($P_{K^+}/P_{Na^+} \approx 1,000$), with energetic barrier differences to these ions reported to be about 6.6 kcal mol⁻¹ for KcsA²⁵.

Diffusion of Cl^- ions across the constriction-site energy barriers was relatively rare on the timescale of our equilibrium molecular dynamics simulations, which necessitated enhanced sampling methods to construct robust Cl^- PMF calculations (see Methods, Extended Data Fig. 8). Nevertheless, a few Cl^- entry events were observed in our simulation data for Cx50 and Cx46, and in these cases, Cl^- ions appear to co-migrate across the high energy barrier of the constriction site alongside a K^+ counter ion. It is possible that similar mechanisms involving ionic-charge neutralization enable cation-preferring gap junction channels to permit passage of negatively charged signalling molecules (for example, cAMP and Ins(1,4,5)P₃). However, it is difficult to provide a general mechanism for selectivity, as conductance properties for ions do not always correlate well with conductance properties for larger molecules^{41,42}.

The analysis above supports models that propose that substrate selectivity and conductance properties of gap junctions are established by complex mechanisms involving both steric aperture and the unique pattern of electrostatic features contributed by isoform-specific amino acid composition^{42,43}. In this way, Cx46/50 heteromeric or heterotypic channels confer distinct conductance properties of potential functional significance. For example, rectification observed in Cx46/50 heterotypic channels can be explained by the resulting asymmetric free-energy landscape (Extended Data Fig. 8) induced by the uneven distribution of fixed charges^{30,40}. Cx46/50 heteromeric assemblies also produced unique K^+ – Cl^- PMF profiles, with peak barriers that were intermediate to their homomeric counterparts (Extended Data Fig. 8), supporting observations made from single-channel measurement^{29,44}.

Additional fine tuning of gap junction permeation properties may be achieved through co-translational and/or post-translational modification of pore-lining residues²². In our studies, N-terminal acetylation was found to enhance the cation-to-anion specificity of Cx50, Cx46 and Cx26 intercellular channels (Extended Data Fig. 8). Although N-terminal acetylation is irreversible, the effect of this co-translational modification illustrates how other dynamic and reversible charge-modifying post-translational modifications may serve to spatially and temporally modulate the behaviour of intercellular communication.

Open-state stabilization of the NTH domain

Despite general similarities, substantial differences between the Cx46/50 cryo-EM structures and the Cx26 crystal structure are localized to the NTH domain (Fig. 4, Extended Data Fig. 7). The connexin NTH domain contributes to ion selectivity and ‘fast’ trans-junctional voltage gating that is common to all connexin isoforms⁴⁵. The NTH folds into the cytoplasmic vestibule, where it forms the constriction site, and is well-positioned to function as a selectivity filter or gating domain (Figs. 1d, 4a).

In the proposed open-state conformation of Cx46/50 described here, the NTH domain adopts a regular amphipathic α -helix and ordered loop connecting to TM1 (Fig. 4a–c). The hydrophobic face is established by a set of aromatic and hydrophobic residues that are conserved across various connexin isoforms (W4, L7, I10, L11 and V14 in Cx46 and Cx50) (Fig. 4b). These anchoring sites pack against the pore-lining helices (TM1–TM2), and along the interface of neighbouring subunits. Despite sequence conservation at these sites, the NTH domain

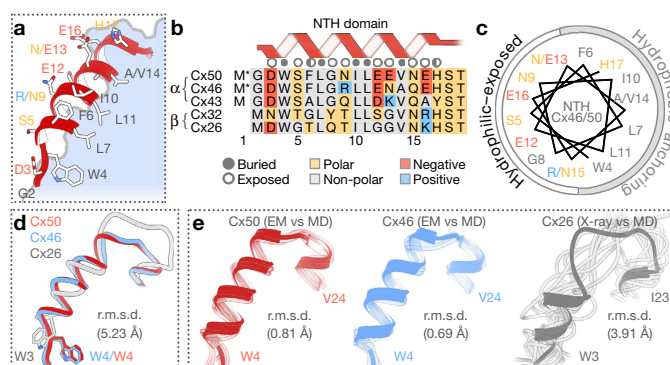


Fig. 4 | Open-state stabilization of the Cx46/50 NTH domain.

a, Magnified view of the Cx46/50 NTH domain. **b**, Sequence alignment of the NTH domain from representative α - and β -family connexins. Residues are annotated as buried, exposed or partially exposed. *, co-translational removal of M1, as confirmed by MS/MS. **c**, Helical wheel representation of the Cx46/50 NTH domain. In **a**–**c**, amino acids are labelled with a single letter for residues that are the same for both Cx46 and Cx50, and separately in the format (Cx46/Cx50) where they differ, and coloured by chemical properties (grey, hydrophobic; blue, positively charged; red, negatively charged; and yellow, hydrophilic). **d**, Overlay of NTH domains from experimental structures of Cx50 (red), Cx46 (blue) and Cx26 (grey; PDB 2ZW3¹⁰), after super-positioning of TM1–TM4, EC1 and EC2 domains. **e**, Superposition of NTH domains of each monomer captured from the molecular dynamics simulation (MD, faded tube), and aligned against the initial starting structures (X-ray or cryo-EM (EM)), Cx50ac (left), Cx46ac (centre) and Cx26ac (right), displayed as ribbons. C α r.m.s.d. of the NTH domains are shown in **d** and **e**.

modelled in the crystal structure of Cx26 is in a distinctively different conformation and overall arrangement with respect to the trans-membrane domains compared to Cx46/50 (C α r.m.s.d. = 5.2 Å, after alignment of the transmembrane and extracellular domains) (Fig. 4d, Extended Data Fig. 7e). In Cx26, the NTH domain and loop connecting TM1 is less regular, and with the exception of W3 (Cx26 numbering) the conserved hydrophobic residues are modelled towards the solvent¹⁰.

We propose that the network of hydrophobic anchoring observed in the cryo-EM structure of Cx46/50 supports a stabilized open-state conformation. Accordingly, analyses of our molecular dynamics simulations show that the NTH domains of Cx50 and Cx46 are conformationally stable in both acetylated and non-acetylated states, with only small-amplitude backbone fluctuations (root mean square fluctuation (r.m.s.f.) ≈ 1.0 –1.2 Å) (Fig. 4e, Extended Data Fig. 7). By contrast, the NTH domain of Cx26 (and Cx26ac) becomes rapidly disordered (that is, unfolded), and remains conformationally dynamic throughout the production phase of our molecular dynamics simulations (using Protein Data Bank (PDB) code 2ZW3¹⁰ as the starting structure; Fig. 4e, Extended Data Fig. 7). The dynamical behaviour of the Cx26 NTH domain is consistent with previous molecular dynamics studies^{35,46,47}. The functional significance of the differences in NTH domain structure and dynamic stability is currently unclear. Indeed, instability of the Cx26 NTH domain may be an intrinsic feature. In a more recent X-ray crystallographic study of Cx26, the NTH domain was completely unresolved, presumably owing to local disorder¹¹; however, potential effects of the conditions required for crystallization cannot be ruled out.

The amphipathic nature of the Cx46/50 NTH positions hydrophilic residues implicated in voltage sensing and ion selectivity at the solvent-exposed face, forming the cytoplasmic vestibule⁴⁵ (Fig. 4a–c). A network of hydrogen-bond interactions appears to contribute to the precise localization of some of these key residues, including the site of N-terminal acetylation. The carbonyl group at the acetylated-G2 site appears to be oriented—at least transiently—through hydrogen bonding to the indole ring of W4 in the same subunit, whereas in the non-acetylated state, G2 forms a transient intermolecular ion pair with D3 of a neighbouring subunit (Extended Data Fig. 10). The side chain of D3 is oriented by a relatively stable intramolecular hydrogen

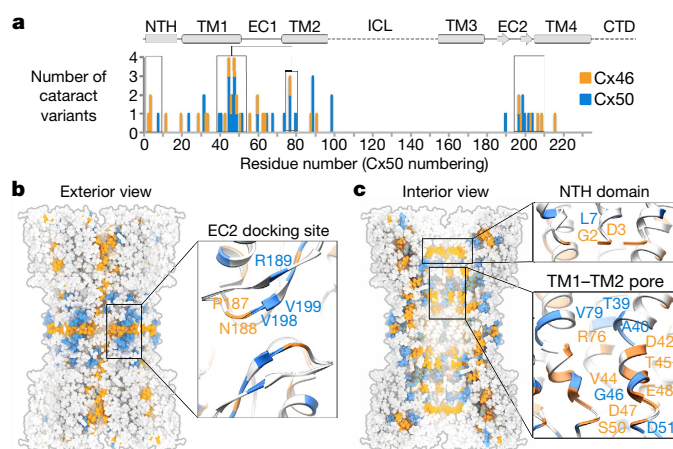


Fig. 5 | Mutational hot spots in Cx46 and Cx50 linked to congenital cataracts. **a**, Histogram of genetic variants of Cx46 and Cx50 linked to hereditary (congenital) cataracts. Sites of Cx46 (orange) and Cx50 (blue) mutations are overlaid using Cx50 amino acid numbering, with secondary structure and domain elements indicated. Hot spots, regions of high genetic variation, within the NTH, TM1, TM2 and EC2 domains are boxed. **b**, **c**, Exterior (**b**) and interior (**c**) view of the Cx46/50 gap junction channel with cataract mutation sites mapped for Cx46 (orange) and Cx50 (blue), with magnified views of the EC2 domain (**b**, inset), NTH domain (**c**, top inset) and TM1–TM2 pore-lining helices (**c**, bottom inset), with representative mutation sites labelled.

bond with the hydroxyl of S5, in both acetylated and non-acetylated forms of both channels (Extended Data Fig. 10). D3 (in Cx46/50, D2 in Cx26) has been identified as a critical site for establishing polarity and/or sensitivity to trans-junctional voltage^{48–50}. The precise spatial orientation imposed by these interactions may contribute to the strict conservation at this site, as replacement of D3 to a similarly negatively charged residue—glutamate—results in significant perturbation to gating properties, conductance and free energy of the open–closed state in Cx50 gap junctions⁵¹.

Mutational hot spots linked to hereditary cataracts

Cx46/50 gap junctions have a critical role in maintaining the transparency of the eye lens by establishing a pathway for water, ion and nutrient circulation and removal of metabolic waste in this avascular organ¹². Consequently, a variety of human genetic variations in Cx50 and Cx46 have been linked to hereditary cataract formation⁵². Age-related cataracts are currently incurable (except by surgery) and remain the leading cause of blindness in the world⁵³. The rarer congenital forms of this disease have been linked to genetic mutation of various lens proteins, including Cx46/50—offering critical insight into the mechanisms of maintaining lens transparency throughout life⁵⁴. We mapped 46 mutation sites in Cx46/50, currently reported on the Cat-Map database⁵⁴, that are linked to congenital cataracts (Fig. 5a). This analysis suggests explanations for many of the disease-causing effects induced by these polymorphisms, as the mutational hot spots localize to functionally important regions of the Cx46/50 gap junction structure. These include a cluster of residues localized within the EC2 docking site—for example, Cx46(N188T/I) and Cx50(R189Q/W)—in regions that deviate markedly from the Cx26 structure, such as the NTH gating or selectivity domain—such as Cx46(G2D) and Cx46(D3Y/H), and Cx50(L7P)—and several sites localized to the TM1–TM2 pore-lining helix that form an interaction network with the NTH domain, of which mutation is expected to affect the permeation pathway or folding and stability within these regions (Fig. 5b, c). The localization of disease-causing mutations underscores the functional significance of these core structural–functional elements, and the importance of proper cell-to-cell communication through Cx46/50 gap junctions for the maintenance of lens transparency. The ability of cryo-EM to provide high-resolution structural information on gap junctions may finally

enable detailed mechanistic investigation of these disease-causing mutations in Cx46/50 and in other isoforms responsible for a diverse range of connexinopathies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0786-7>.

Received: 11 September 2017; Accepted: 29 October 2018;

Published online 12 December 2018.

- Goodenough, D. A. & Paul, D. L. Gap junctions. *Cold Spring Harb. Perspect. Biol.* **1**, a002576 (2009).
- Delmar, M. et al. Connexins and disease. *Cold Spring Harb. Perspect. Biol.* **10**, a029348 (2018).
- Garcia, I. E. et al. Connexinopathies: a structural and functional glimpse. *BMC Cell Biol.* **17** (Suppl 1), 17 (2016).
- Aasen, T., Mesnil, M., Naus, C. C., Lampe, P. D. & Laird, D. W. Gap junctions and cancer: communicating for 50 years. *Nat. Rev. Cancer* **16**, 775–788 (2016).
- Sosinsky, G. E. & Nicholson, B. J. Structural organization of gap junction channels. *Biochim. Biophys. Acta* **1711**, 99–125 (2005).
- Sohl, G. & Willecke, K. Gap junctions and the connexin protein family. *Cardiovasc. Res.* **62**, 228–232 (2004).
- Cottrell, G. T. & Burt, J. M. Functional consequences of heterogeneous gap junction channel formation and its influence in health and disease. *Biochim. Biophys. Acta* **1711**, 126–141 (2005).
- Beyer, E. C. & Berthoud, V. M. Gap junction structure: unraveled, but not fully revealed. *F1000 Res.* **6**, 568 (2017).
- Grosely, R. & Sorgen, P. L. A history of gap junction structure: hexagonal arrays to atomic resolution. *Cell Commun. Adhes.* **20**, 11–20 (2013).
- Maeda, S. et al. Structure of the connexin 26 gap junction channel at 3.5 Å resolution. *Nature* **458**, 597–602 (2009).
- Bennett, B. C. et al. An electrostatic mechanism for Ca²⁺-mediated regulation of gap junction channels. *Nat. Commun.* **7**, 8770 (2016).
- Mathias, R. T., White, T. W. & Gong, X. Lens gap junctions in growth, differentiation, and homeostasis. *Physiol. Rev.* **90**, 179–206 (2010).
- König, N. & Zampighi, G. A. Purification of bovine lens cell-to-cell channels composed of connexin44 and connexin50. *J. Cell Sci.* **108**, 3091–3098 (1995).
- Jiang, J. X. & Goodenough, D. A. Heteromeric connexons in lens gap junction channels. *Proc. Natl Acad. Sci. USA* **93**, 1287–1291 (1996).
- Shearer, D., Ens, W., Standing, K. & Valdimarsson, G. Posttranslational modifications in lens fiber connexins identified by off-line-HPLC MALDI-quadrupole time-of-flight mass spectrometry. *Invest. Ophthalmol. Vis. Sci.* **49**, 1553–1562 (2008).
- Wang, Z. & Schey, K. L. Phosphorylation and truncation sites of bovine lens connexin 46 and connexin 50. *Exp. Eye Res.* **89**, 898–904 (2009).
- Kwon, T. et al. Molecular dynamics simulations of the Cx26 hemichannel: insights into voltage-dependent loop-gating. *Biophys. J.* **102**, 1341–1351 (2012).
- Bai, D. Structural analysis of key gap junction domains—lessons from genome data and disease-linked mutants. *Semin. Cell Dev. Biol.* **50**, 74–82 (2016).
- Koval, M., Molina, S. A. & Burt, J. M. Mix and match: investigating heteromeric and heterotypic gap junction channels in model systems and native tissues. *FEBS Lett.* **588**, 1193–1204 (2014).
- Unger, V. M., Kumar, N. M., Gilula, N. B. & Yeager, M. Three-dimensional structure of a recombinant gap junction membrane channel. *Science* **283**, 1176–1180 (1999).
- Fleishman, S. J., Unger, V. M., Yeager, M. & Ben-Tal, N. A C^α model for the transmembrane alpha helices of gap junction intercellular channels. *Mol. Cell* **15**, 879–888 (2004).
- Gong, X. Q. & Nicholson, B. J. Size selectivity between gap junction channels composed of different connexins. *Cell Commun. Adhes.* **8**, 187–192 (2001).
- Goldberg, G. S., Valiunas, V. & Brink, P. R. Selective permeability of gap junction channels. *Biochim. Biophys. Acta* **1662**, 96–101 (2004).
- Varland, S., Osberg, C. & Arnesen, T. N-terminal modifications of cellular proteins: the enzymes involved, their substrate specificities and biological effects. *Proteomics* **15**, 2385–2401 (2015).
- Berneche, S. & Roux, B. Energetics of ion conduction through the K⁺ channel. *Nature* **414**, 73–77 (2001).
- Corry, B. & Thomas, M. Mechanism of ion permeation and selectivity in a voltage gated sodium channel. *J. Am. Chem. Soc.* **134**, 1840–1846 (2012).
- Ulmschneider, M. B. et al. Molecular dynamics of ion transport through the open conformation of a bacterial voltage-gated sodium channel. *Proc. Natl Acad. Sci. USA* **110**, 6364–6369 (2013).
- Srinivas, M. et al. Voltage dependence of macroscopic and unitary currents of gap junction channels formed by mouse connexin50 expressed in rat neuroblastoma cells. *J. Physiol.* **517**, 673–689 (1999).
- Hopperstad, M. G., Srinivas, M. & Spray, D. C. Properties of gap junction channels formed by Cx46 alone and in combination with Cx50. *Biophys. J.* **79**, 1954–1966 (2000).
- Oh, S., Rubin, J. B., Bennett, M. V., Verselis, V. K. & Bargiello, T. A. Molecular determinants of electrical rectification of single channel conductance in gap junctions formed by connexins 26 and 32. *J. Gen. Physiol.* **114**, 339–364 (1999).

31. Tong, X. et al. The first extracellular domain plays an important role in unitary channel conductance of Cx50 gap junction channels. *PLoS ONE* **10**, e0143876 (2015).
32. Trexler, E. B., Bukauskas, F. F., Kronengold, J., Bargiello, T. A. & Verselis, V. K. The first extracellular loop domain is a major determinant of charge selectivity in connexin46 channels. *Biophys. J.* **79**, 3036–3051 (2000).
33. Oh, S., Verselis, V. K. & Bargiello, T. A. Charges dispersed over the permeation pathway determine the charge selectivity and conductance of a Cx32 chimeric hemichannel. *J. Physiol.* **586**, 2445–2461 (2008).
34. Kronengold, J., Trexler, E. B., Bukauskas, F. F., Bargiello, T. A. & Verselis, V. K. Pore-lining residues identified by single channel SCAM studies in Cx46 hemichannels. *Cell Commun. Adhes.* **10**, 193–199 (2003).
35. Zonta, F., Polles, G., Zanotti, G. & Mammano, F. Permeation pathway of homomeric connexin 26 and connexin 30 channels investigated by molecular dynamics. *J. Biomol. Struct. Dyn.* **29**, 985–998 (2012).
36. Lopez, W. et al. Mechanism of gating by calcium in connexin hemichannels. *Proc. Natl Acad. Sci. USA* **113**, E7986–E7995 (2016).
37. Harris, A. L. & Contreras, J. E. Motifs in the permeation pathway of connexin channels mediate voltage and Ca^{2+} sensing. *Front. Physiol.* **5**, 113 (2014).
38. Srinivas, M., Calderon, D. P., Kronengold, J. & Verselis, V. K. Regulation of connexin hemichannels by monovalent cations. *J. Gen. Physiol.* **127**, 67–75 (2006).
39. Trexler, E. B., Bennett, M. V., Bargiello, T. A. & Verselis, V. K. Voltage gating and permeation in a gap junction hemichannel. *Proc. Natl Acad. Sci. USA* **93**, 5836–5841 (1996).
40. Suchyna, T. M. et al. Different ionic selectivities for connexins 26 and 32 produce rectifying gap junction channels. *Biophys. J.* **77**, 2968–2987 (1999).
41. Veenstra, R. D. Size and selectivity of gap junction channels formed from different connexins. *J. Bioenerg. Biomembr.* **28**, 327–337 (1996).
42. Veenstra, R. D. et al. Selectivity of connexin-specific gap junctions does not correlate with channel conductance. *Circ. Res.* **77**, 1156–1165 (1995).
43. Nicholson, B. J. et al. The molecular basis of selective permeability of connexins is complex and includes both size and charge. *Braz. J. Med. Biol. Res.* **33**, 369–378 (2000).
44. Ebihara, L., Xu, X., Oberti, C., Beyer, E. C. & Berthoud, V. M. Co-expression of lens fiber connexins modifies hemi-gap-junctional channel behavior. *Biophys. J.* **76**, 198–206 (1999).
45. Xin, L. & Bai, D. Functional roles of the amino terminal domain in determining biophysical properties of Cx50 gap junction channels. *Front. Physiol.* **4**, 373 (2013).
46. Luo, Y., Rossi, A. R. & Harris, A. L. Computational studies of molecular permeation through connexin26 channels. *Biophys. J.* **110**, 584–599 (2016).
47. Kwon, T., Harris, A. L., Rossi, A. & Bargiello, T. A. Molecular dynamics simulations of the Cx26 hemichannel: evaluation of structural models with Brownian dynamics. *J. Gen. Physiol.* **138**, 475–493 (2011).
48. Verselis, V. K., Ginter, C. S. & Bargiello, T. A. Opposite voltage gating polarities of two closely related connexins. *Nature* **368**, 348–351 (1994).
49. Peracchia, C. & Peracchia, L. L. Inversion of both gating polarity and CO_2 sensitivity of voltage gating with D3N mutation of Cx50. *Am. J. Physiol. Cell Physiol.* **288**, C1381–C1389 (2005).
50. Srinivas, M., Kronengold, J., Bukauskas, F. F., Bargiello, T. A. & Verselis, V. K. Correlative studies of gating in Cx46 and Cx50 hemichannels and gap junction channels. *Biophys. J.* **88**, 1725–1739 (2005).
51. Xin, L., Nakagawa, S., Tsukihara, T. & Bai, D. Aspartic acid residue D3 critically determines Cx50 gap junction channel transjunctional voltage-dependent gating and unitary conductance. *Biophys. J.* **102**, 1022–1031 (2012).
52. Beyer, E. C., Ebihara, L. & Berthoud, V. M. Connexin mutants and cataracts. *Front. Pharmacol.* **4**, 43 (2013).
53. Pascolini, D. & Mariotti, S. P. Global estimates of visual impairment: 2010. *Br. J. Ophthalmol.* **96**, 614–618 (2012).
54. Shiels, A. & Hejtmancik, J. F. Mutations and mechanisms in congenital and age-related cataracts. *Exp. Eye Res.* **156**, 95–102 (2017).

Acknowledgements We thank T. Gonen for early support of this work; T. White, L. David, U. Adhikari and B. Mostofian for helpful discussions; the staff at the OHSU Multiscale Microscopy Core and Advanced Computing Center, and W. Garrick (PSU) for their assistance and training. C.V.R. and D.S.C. are supported by funding from the European Research Council (No. 695511-ENABLE). D.M.Z. and C.C.Y. are supported by the Center for Spatial Systems Biomedicine at OHSU. S.L.R. is supported by the Medical Research Foundation of Oregon and the National Institutes of Health (R35-GM124779).

Author contributions J.B.M. and B.G.H. contributed equally. J.B.M. and C.C.Y. collected the cryo-EM datasets. J.B.M. performed image processing and atomic modelling of Cx46/50. S.E.O. performed protein purification and negative-stain electron microscopy studies. B.G.H. conducted and analysed the molecular dynamics simulations. D.S.C. conducted the cross-linking studies and MS/MS analysis. C.V.R. contributed to the experimental design of MS/MS studies. D.M.Z. contributed to the experimental design and statistical analysis of the molecular dynamics simulations. All authors contributed to manuscript preparation. S.L.R. provided overall guidance to the design and execution of this work.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0786-7>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0786-7>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to S.L.R.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Cx46/50 purification and amphipol reconstitution. Lamb eyes were obtained from the Wolverine Packers slaughterhouse (Detroit), and the lenses were removed using a surgical blade and stored at -86°C . Gap junction intercellular channels were isolated from the core lens fibre tissue, containing C-terminal truncation variants of Cx46 and Cx50 (also known as MP38)^{16,55–57} (Extended Data Fig. 1). Details of the purification procedure are provided below.

Core lens-fibre cell tissue was dissected from cortical tissue using a surgical blade, and stripped membranes were prepared as described^{58–60}. Total protein concentration was determined by BCA (Pierce) and membranes were stored at -86°C suspended in storage buffer containing 10 mM Tris pH8.0, 2 mM EDTA, 2 mM EGTA, at a total protein concentration of $\sim 2\text{ mg ml}^{-1}$. Stripped membranes were thawed from -86°C and solubilized in 10 mM Tris pH8.0, 2 mM EDTA, 2 mM EGTA, 1% (w/v) *n*-decyl- β -D-maltoside (DM) for 30 min at 37°C . Unsolubilized debris was cleared by ultracentrifugation at $150,000g$ for 30 min at 4°C . The solubilized fraction was applied to an anion-exchange chromatography column (UnoQ, BioRad) equilibrated with buffer A (10 mM Tris pH 8.0, 2 mM EDTA, 2 mM EGTA, 0.3% DM (w/v)). Protein was eluted with buffer B, which additionally contained 500 mM NaCl. Elution peaks containing Cx46/50, as determined by SDS–PAGE, were pooled and applied to a size-exclusion chromatography (SEC) column (ENC650, BioRad) equilibrated with SEC buffer (20 mM HEPES, 150 mM NaCl, 2 mM EDTA, 2 mM EGTA and 0.3% DM (w/v)). Peak fractions containing purified Cx46/50 were pooled and protein concentration was determined by UV absorbance. All chromatography steps were performed at 4°C . The presence of both Cx46 and Cx50 was confirmed by western blot analysis using polyclonal antibodies directed against the N-terminal domain of Cx46 (AP11570PU-N, Acris) and C-terminal domain of Cx50 (sc-50432, Santa Cruz) (Extended Data Fig. 1a) and by mass spectrometry analysis, described below.

Purified Cx46/50 was exchanged from DM to amphipol A8-35 (Anatrace), as follows. Amphipol was added to freshly purified protein in an 5:1 amphipol:protein (w/w) ratio using a stock solution prepared at 5% (w/v) in water. This mixture was incubated for 2.5 h at 4°C with rotation. Detergent was then removed by application of SM-2 Biobeads (BioRad) at a ratio of 30:1 (w/w) beads:detergent. Biobeads were incubated overnight at 4°C with rotation. Biobeads were then removed by running samples over a Polyrep column (BioRad) that had been washed with detergent-free SEC buffer (20 mM HEPES pH 7.4, 150 mM NaCl, 2 mM EDTA, 2 mM EGTA). Samples were further cleared by ultra-centrifugation at $150,000g$ for 20 min at 4°C . The clarified sample was then applied to an SEC column (ENC650, BioRad) equilibrated with detergent-free SEC buffer to remove excess amphipol. Peak fractions corresponding to amphipol-stabilized Cx46/50 were pooled and concentrated for single-particle electron microscopy studies (Extended Data Fig. 1b). Final protein concentration was determined by UV absorbance at 280 nm.

Chemical cross-linking and mass spectrometry. Cx46/50 was prepared for cross-linking and mass spectrometry analysis to confirm heteromeric assembly of the two lens isoforms (Extended Data Fig. 1d–f). Cross-linking was carried out using the amine-to-amine cross-linking reagents bis(sulfosuccinimidyl)suberate (BS3; Thermo Fisher Scientific) and disuccinimidyl suberate (DSS H12:D12; Creative Molecules). For BS3 cross-linking, $0.5\text{ }\mu\text{l}$ of 12.5 mM BS3 dissolved in deuterium-depleted water (DDW) were added to $10\text{ }\mu\text{l}$ of purified $20\text{ }\mu\text{M}$ Cx46/50 complexes and incubated for 2 h on ice. For DSS cross-linking, $0.5\text{ }\mu\text{l}$ of a 1:1 mixture of 25 mM non-deuterated (d0) and deuterated (d12) DSS cross-linker dissolved in DMSO was added to $10\text{ }\mu\text{l}$ purified $20\text{ }\mu\text{M}$ Cx46/50 complexes. A control sample, in which only DMSO or DDW were added to the proteins, was also prepared. DSS cross-linked samples were incubated at room temperature for 2 h in a thermomixer at 300 revolutions per min (r.p.m.), whereas BS3 cross-linked samples were incubated for 2 h on ice. The cross-linking reaction was quenched by adding Tris pH7.4 at a final concentration of 100 mM for 15 min at room temperature in a thermomixer at 300 r.p.m.

The quenched reaction mixtures were separated on a NuPAGE gel (Thermo Fisher Scientific) and protein bands stained with InstantBlue (Expedeon). Cross-linked protein bands were excised and digested with trypsin (Promega) as described⁶¹. Peptides were resuspended in 0.1% formic acid and separated on an Ultimate 3000 UHPLC system (Thermo Fisher Scientific) and electrosprayed directly into a QExactive mass spectrometer (Thermo Fisher Scientific) through an EASY-Spray nano-electrospray ion source (Thermo Fisher Scientific). The peptides were trapped on a C18 PepMap100 pre-column ($300\text{ }\mu\text{m i.d.} \times 5\text{ mm}$, 100 Å, Thermo Fisher Scientific) using solvent A (0.1% formic acid in water) at a pressure of 500 bar. The peptides were separated on an in-house packed analytical column ($75\text{ }\mu\text{m}$ internal diameter packed with ReproSil-Pur 120 C18-AQ, $1.9\text{ }\mu\text{m}$, 120 Å, Dr. Maisch) using a gradient (length: 50 min, 15% to 38% for 30 min followed by 38% to 58% solvent B (0.1% formic acid in acetonitrile, flow rate: 200 nl/min) for 15 min. Raw data were acquired on the mass spectrometer in a data-dependent mode (DDA). Full-scan mass spectra were acquired in the Orbitrap (scan range 350–2000 m/z , resolution 70,000, AGC target 3×10^6 , maximum injection time 50 ms).

After the mass spectrometry scans, the 10 most intense peaks were selected for HCD fragmentation at 30% of normalized collision energy. HCD spectra were also acquired in the Orbitrap (resolution 17500, AGC target 5×10^4 , maximum injection time 120 ms) with first fixed mass at 180 m/z . Charge exclusion was selected for 1+ and 2+ ions. The dynamic exclusion set to 5 s. Cross-linking identification and analysis was done using pLink⁶² and Xcalibur 2.2 (Thermo Scientific). All peptides were manually validated.

For identification of proteins and post-translational modifications, protein bands were excised from the gel and processed as described above. Mass spectrometry analysis was carried out similarly with a gradient of 15–38% for 30 min and the Orbitrap set to 350–1,500 m/z . Charge exclusion was selected for 1+ and unassigned ions, dynamic exclusion was set to 5 s. PTM identification was done using the MASCOT Daemon client program.

Negative-stain electron microscopy. Amphipol-stabilized Cx46/50 was prepared for negative-stain electron microscopy as described^{58,63}. In brief, $3\text{ }\mu\text{l}$ sample ($\sim 0.02\text{ mg ml}^{-1}$) was applied to a glow-discharged continuous carbon coated electron microscopy specimen grid (Ted Pella), blotted with filter paper and washed two times with detergent-free SEC buffer. The specimen was then stained with freshly prepared 0.75% (w/v) uranyl formate (SPI-Chem).

Negatively stained specimens were visualized on a 120 kV TEM (iCorr, FEI) at a nominal magnification of $49,000\times$ at the specimen level (Extended Data Fig. 1c). Digital micrographs were recorded on a $2k \times 2k$ CCD camera (FEI Eagle) with a calibrated pixel size of 4.37 Å . A total of 75 micrographs were collected. Contrast transfer function (CTF) parameters were determined in EMAN2⁶⁴ and micrographs free of significant astigmatism and drift were selected based on Thon rings in the power spectra. A total of 5,330 particles were hand-selected in EMAN2 and extracted with a box size of 84×84 pixels. Reference-free 2D class averages were generated using CTF-corrected (phase-flipped) images without applied symmetry (Extended Data Fig. 1c). A subset of 3,952 ‘good’ particles was selected following multiple rounds of 2D classification, and an initial model was generated de novo in EMAN2 using a subset of 12 class averages as input. This model was refined against the good particle image dataset in EMAN2 with applied D6 symmetry to a final resolution of $\sim 20\text{ Å}$ (Extended Data Fig. 2).

Cryo-EM data collection, image processing and 3D reconstruction. Samples were prepared for cryo-EM by applying $5\text{ }\mu\text{l}$ of amphipol-stabilized Cx46/50 (2.35 mg ml^{-1}) to a glow-discharged holey carbon grid (Quantifoil R1.2/1.3) for 10 s. The grid was blotted for 4.0 s and plunge-frozen in liquid ethane using a Vitrobot (FEI) at 100% humidity and stored under liquid nitrogen.

Cryo-EM specimen grids were imaged on a Titan Krios (FEI) operated at 300 kV. Image stacks were recorded using a K2 summit direct electron detector (Gatan) in counting mode with a super-resolution pixel size of 0.665 Å/pix . The dose rate was $3.2\text{ electrons pixel}^{-1}\text{ s}^{-1}$, with 4 frames s^{-1} collected for a total exposure time of 10 s. A Gatan energy filter with a slit width of 30 eV was used during data collection. An initial dataset of 1,104 micrographs (dataset 1) was obtained by automated data collected using SerialEM⁶⁵, with nominal defocus values from 1.25 to $2.5\text{ }\mu\text{m}$.

Drift correction and dose weighting was performed using MotionCor2⁶⁶ and CTF correction was performed using GCTF⁶⁷. Then 261,206 particles were picked from dataset 1 using DoGPicker⁶⁸. Particles were extracted with $2\times$ binning (resulting in a pixel size of 1.3 Å per pixel). Five rounds of 2D classification in Relion 2.0⁶⁹ left 53,791 good particles. These particles were then subjected to 3D classification in Relion with four classes and no imposed symmetry. The most populated class contained 33,967 particles. These particles were unbinned and another round of 3D classification was performed, reducing the population of particles to 30,128. 3D auto-refinement was then performed on this set of particles with D6 symmetry imposed. After masking and post-processing in Relion, the final map had a resolution of 3.4 Å by gold-standard Fourier shell correlation (FSC) (Extended Data Figs. 2, 3a).

An additional 1,093-micrograph dataset was collected and processed as above. Particles were picked from this set with DoGPicker, $2\times$ binned and pooled with the original set of 261,206 particles for a total of 398,066 particles (dataset 2). A set of 66,480 good particles was obtained after five rounds of 2D classification. These particles were subjected to 3D classification with four classes and no imposed symmetry. The most populated class contained 55,475 particles. This dataset was further culled by removing particles extracted from micrographs with Thon rings that did not extend beyond 3.5 Å , resulting in a final dataset of 44,547 particles. These particle images were subjected to 3D auto-refinement in Relion with D6 symmetry, resulting in an overall 3.5 Å resolution 3D reconstruction after post-processing as judged by gold-standard FSC (Extended Data Fig. 3b). Local resolution analysis using BlocRes⁷⁰ showed the 3.4 Å map possessed exceptionally high-resolution features within the central regions of the structure, whereas the 3.5 Å reconstruction contained more uniformly defined features throughout the density map, consistent with visual inspection (Extended Data Fig. 3c, d). An overview of cryo-EM data collection and 3D refinement statistics is provided in Extended Data Fig. 2.

Cx46/50 symmetry analysis. In an attempt to uncover a specific pattern(s) of Cx46/50 heteromeric/heterotypic co-assembly, 3D auto-refinement was also pursued in Relion using C1, C3, C6 and D3 symmetries, using the final 30,128-particle (dataset 1) and 3.4 Å map (filtered to 15 Å) as input. These refinements converged to 4.1 Å (C1), 3.9 Å (C3) and 3.7 Å (D3 and C6). Examination of the resulting maps provided no indication that the Cx46 and Cx50 subunits were being separately resolved (Extended Data Figs. 4, 5). Further attempts were performed using 3D classification in Relion with C3, C6 and D3 symmetry, using the larger 55,475-particle set (which had already been subjected to one round of 3D classification with no imposed symmetry). The initial model was the 3.4 Å map filtered to 25 Å. No resolution limit was enforced, and classification was attempted with and without image alignment. Some classifications converged to a single class, whereas others maintained a more even distribution of particles throughout 3D classification. 3D auto-refine was attempted with the most populated class from each attempted symmetry group. C3 symmetry refined to 3.9 Å from a set of 47,074 particles; C6 symmetry refined to 3.8 Å from a set of 38,404 particles; D3 symmetry refined to 4.2 Å from a set of 16,520 particles. Inspection of the resulting maps provided no indication that isoform-specific features were being separately resolved into any specific symmetric arrangements (not shown). Finally, focused refinement strategies with signal subtraction were also explored using Relion, by masking a single hemichannel or just a single subunit. However, these procedures did not produce isoform-specific features, or improved results compared to the D6-symmetrized maps.

As we were unable to identify a specific pattern of co-assembly for the Cx46/50 dodecameric channel, all further analysis and model building was performed using the 3D maps generated with imposed D6 symmetry. Both pre-processed and post-processed maps and associated masks generated from datasets 1 and 2 have been deposited in the Electron Microscopy Data Bank under accession code EMD-9116.

Atomic modelling, refinement and validation. The post-processed maps obtained with D6 symmetry were used to build and stereochemically refine atomic models for both Cx46 and Cx50, following similar procedures. An initial C_α model was generated using the available crystal structure of Cx26 (PDB 2ZW3¹⁰) and placed into the post-processed 3.4 Å density map using rigid-body fitting. Starting from this template, all atom models of Cx46 and Cx50 were built separately into the cryo-EM density using COOT⁷¹. Disulfide bonds were modelled for Cx50 (C54–C201, C61–C195 and C65–C190) and Cx46 (C54–C189, C61–C183 and C65–C178). Models were subjected to real-space refinement in Phenix⁷² with non-crystallographic symmetry (D6-symmetry) and secondary structure restraints imposed. Successive rounds of modelling and refinement were conducted until refinement statistics converged, as judged by Molprobity⁷³ (Extended Data Fig. 2). The FSC of the model versus map dropped below 0.5 at 3.4 Å (dataset 1) and 3.5 Å (dataset 2) for both Cx46 and Cx50, judged by the output of Phenix real-space refine (Extended Data Fig. 3a, b). The NTH domain of Cx46 and Cx50 (residues 2–20) were further refined using the post-processed 3.5 Å density map (dataset 2), as this region of the map was more well-defined compared to the original 3.4 Å map. Over areas of the density maps where the sequence of Cx46 and Cx50 are identical or similar (80% identical and 8% similar) both models fit well into the D6-symmetrized map, and these regions tend to display well-resolved side-chain density. Over regions in which the sequence of Cx46 and Cx50 differs, side-chain density is sometimes weaker. This observation is possibly due to the imposed D6 symmetry averaging the density of two different side chains in these areas, or relative flexibility as many of these residues contain solvent-exposed side chains. In these areas of difference, where electron microscopy density is observed, both Cx46 and Cx50 can be fit into the density equally well (Extended Data Figs. 4, 5). Fit of the models to the cryo-EM density map were assessed quantitatively by local resolution analysis using BlocRes⁷⁰, comparing the calculated maps of Cx50 and Cx46 atomic models to the 3.4 Å experimental cryo-EM map (Extended Data Fig. 3). This analysis was tabulated by assigning each residue a range of resolution values corresponding to the output of this analysis, including the alpha carbon and extending to the end of the side chain (Supplementary Tables 2, 3).

Completed models of the dodecameric structures—corresponding to residues 2–97, 142–222 (Cx46), and 2–97, 154–234 (Cx50)—have been deposited in the Protein Data Bank with accession numbers 6MHQ and 6MHY, respectively. Additional density is observed for the region of TM2 that extends towards the cytoplasm; however, we did not model this region (corresponding to ~1–2 turns of an α -helix) owing to the lack of identifiable side-chain density. Various heterotypic/heteromeric models of Cx46/50 were generated for analysis by applying appropriate symmetry operations to the monomeric subunits and combined to form a complete gap junction structure. Coulombic surface potentials were calculated and displayed using Chimera⁷⁴.

Molecular dynamics simulations. Visual Molecular Dynamics (VMD) v1.9.3⁷⁵ was used to build systems for the Cx50, Cx46, Cx46/50 heteromeric and heterotypic models, and for Cx26 (PDB 2ZW3)¹⁰. Representative Cx46/50 heteromeric

models (heteromeric models I and II, with C3 or D3 point group symmetry, respectively) were constructed by applying the appropriate symmetry operations to the coordinates of the individual subunits. The Cx46/50 heteromeric and heterotypic channels were run through a steepest descent minimization routine using Phenix⁷² to ensure no clashes were introduced in the preparation of these models. Each system comprised the full dodecameric gap junction, and was prepared in explicit solvent and embedded in two lipid bilayers composed of 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine (POPC), mimicking a cell–cell junction. The Cx26 crystal structure was prepared for molecular dynamics by completing the side chains at residues Lys15, Ser17 and Ser19 and missing protons were added to all amino acids at standard positions. Side chains were protonated according to neutral conditions, and the HSD model was used for all histidine residues. To facilitate comparison to the Cx46/50 models, the Cx26 model was constructed with a Met1 residue added, which was missing in the published crystal structure, but was expected to be present in the protein based on proteomic analysis⁷⁶, as previously described⁴⁷. Disulfide bonds identified in the experimental structures were enforced for Cx50 and Cx46 (as described above), and for Cx26¹⁰ (C53–C180, C60–C174 and C64–C169). Amino acids corresponding to the intracellular loop (ICL) connecting TM2–TM3, and the C-terminal domain (CTD) of Cx50, Cx46 and Cx26 and were not included for molecular dynamics simulation, as experimental data describing the structure of these large domains (~50 residue ICL and ~200 residue CTD in Cx46/50) are missing. The introduced N- and C-terminal residues resulting from the missing ICL segment (Cx46 L97 and L142; Cx50 V97 and L154; and Cx26 G109 and K125) were neutralized. N-terminal acetylation sites were introduced in VMD through an all-atom acetylation patch in the automated PSF-Builder. A complete list of modelled residues for each system is provided in Supplementary Tables 1.

The prepared protein structures were submerged in a hydration shell using Solvate v1.0.1⁷⁷. Water was removed from sections of the channel corresponding to transmembrane domains, based on hydrophobic character and localization of amphipol observed in the experimental cryo-EM data (~20–50 Å from the centre of the channel). The VMD membrane-builder plugin was used to add two POPC bilayers, with dimensions of 152 × 152 Å for Cx46, Cx50 and Cx46/50 models, and 155 × 155 Å for Cx26, and lipids overlapping with protein were removed. The entire system was then placed in a water box with dimensions 150 × 150 × 180 Å for Cx46, Cx50 and Cx46/50 models, and 150 × 150 × 183 Å for Cx26, using VMD's Solvate plugin. The system was neutralized using the Autoionize plugin, then 150 mM KCl and 150 mM NaCl were added to the solvent areas corresponding to intracellular and extracellular regions of the simulation box, respectively (see Fig. 3c). A summary of atoms counts for each system is provided in Supplementary Table 1.

GPU-accelerated nanoscale molecular dynamics v2.12⁷⁸ was used for all classical molecular dynamics simulations, using the CHARMM36 force field^{79,80} for all atoms and TIP3P explicit model for water. Each system was prepared following the same minimization and equilibration protocol as follows. An initial minimization of the lipid tails, with all other atoms fixed, was performed for 1 ns with a 1-fs time step, allowing the tails to 'melt'. Next, the system—including lipids, solvent and ions—was allowed to minimize around the protein, with the protein harmonically constrained for 1 ns. For the Cx46/50 heteromeric/heterotypic and acetylated models, a second minimization step was applied, in which the system was free to minimize with a harmonic constraint on the protein backbone to ensure stable quaternary structure. The entire system was then released from restraints and subjected to all-atom equilibration runs using Langevin thermostat, with a constant temperature of 310 K and constant pressure of 1 atm, with 1 or 2-fs time steps and allowed to proceed for 30 ns (see Supplementary Table 1). Periodic boundary conditions were used to allow for the particle mesh Ewald calculation of electrostatics. Finally, all of the models were continued for a minimum of 50 ns of production. Root mean squared deviations (r.m.s.d.) and root mean squared fluctuations (r.m.s.f.) were calculated using VMD. All three gap junctions approached a steady r.m.s.d. within 20 ns of the equilibration phase (Extended Data Fig. 7a, b). All of these systems maintained an electro-chemical seal to extracellular sodium ions (Na^+) during molecular dynamics simulation (for example, Fig. 3c), validating the stability of intermolecular docking-site interactions and the various heteromeric/heterotypic models generated for analysis.

Calculation of the PMF with respect to K^+ and Cl^- was performed using the fundamental principle of detailed balance via a one-dimensional Markov state model (MSM). Configuration space was subdivided based on a natural coordinate, the channel pore (z axis), and segmented into bins of 4 Å in length. Using a lag-time of 2 ps, a transition matrix was calculated from the trajectories of individual ions within the simulation. The $i \rightarrow j$ transition probability k_{ij} is computed using equation (1):

$$k_{ij} \approx \frac{N_{i,j}}{N_i} \quad (1)$$

in which N_{ij} is the count of transitions during the lag interval and N_i is the count of ions in bin i at the beginning of each lag interval. PMFs were constructed using the principle of detailed balance:

$$P_i^{\text{eq}} k_{i,i+1} = P_{i+1}^{\text{eq}} k_{i+1,i} \quad (2)$$

$$e^{\frac{-\Delta G_{i,i+1}}{RT}} = \frac{P_{i+1}^{\text{eq}}}{P_i^{\text{eq}}} = \frac{k_{i,i+1}}{k_{i+1,i}} \quad (3)$$

$$\text{PMF}(i) = \Delta G(i) = \sum_{n=1}^{i-1} -RT \ln \left(\frac{k_{n,n+1}}{k_{n+1,n}} \right) \quad (4)$$

Here, P_i^{eq} is the equilibrium probabilities for an ion to occupy the respective bin (equation (2)), $\Delta G_{i,i+1}$ is the free-energy difference from bin i to bin $i+1$, R is the gas constant ($1.986 \text{ cal mol}^{-1} \text{ K}^{-1}$), and T is temperature (310 K) (equations (3), (4)). Final PMF values were adjusted so that the values of the bulk solvent were zero. PMF curves in Fig. 3 and Extended Data Fig. 7 were derived by mapping z values to the corresponding bin index i and subsequently smoothed using Microsoft Excel. The detailed-balance (rates-based) approach is justified by the high mobility of ions within the channel pore, and the timescales used for analysis were validated by assessing the convergence of the unsymmetrized data to the symmetrized values presented in Fig. 3d and Extended Data Fig. 8. The results were shown to closely match PMFs constructed by taking the population profile, or average counts, $\langle N_i \rangle$, of the K^+ ions along the channel pore (z axis) and solving: $\Delta G = -RT \ln(\langle N_i \rangle)$ (refs. ^{11,81,82}; Extended Data Fig. 8d).

Because the detailed-balance approach requires only local equilibrium sampling, we were able to apply a distributed seeding protocol to construct PMFs for Cl^- ions. Initial analysis of Cl^- trajectories revealed this ion to be poorly sampled inside the channel pore of the Cx50, Cx46 and Cx26 models, presumably owing to an energetic barrier presented by significant regions of negative coulombic potential for each of these systems (see Fig. 3a). Therefore, a distributed seeding approach was used, in which a single Cl^- ion was randomly introduced (seeded) by replacing a K^+ ion within the pore of the equilibrated channel. These coordinates were energy-minimized and initial velocities were randomized before allowing the simulation to proceed for 10 ns. This procedure was repeated for each system 3–16 times until sufficient sampling was achieved, as determined by monitoring the resulting PMFs. Cl^- PMFs were constructed based on the transition rate (as described above), which is not sensitive to the initial placement of the ion. This seeding approach was validated by showing that the resulting PMF recapitulated the features of a Cl^- PMF obtained for Cx46, where sufficient sampling had been achieved through random diffusion (Extended Data Fig. 8e). Trajectories from these distributed seeding simulations were combined with the production phase data and included in the MSM for calculation of final Cl^- PMFs.

Analysis of hydrogen bonding within the NTH domains of Cx46 and Cx50 models was performed by recording the distance versus time of potential donor-acceptor pairs. The three sets of interactions probed were potential intermolecular hydrogen bonds between D3 and the neighbouring N-terminal G2 residue, the intramolecular hydrogen bonding between D3 and S5, and the intramolecular hydrogen bonds between N-terminal acetyl and W4 for acetylated models of Cx46 and Cx50. To simplify the analysis of hydrogen-bonding interactions involving equivalent rotameric donor-acceptor configurations, heavy atoms were selected for D3 (C_γ) and G2 (N) for analysis (Extended Data Fig. 10). For comparison, equivalent analysis for Cx26 was conducted (between D2 and M1 and T5 of the adjacent subunit). For Cx26, D2 appeared to form intramolecular hydrogen-bonding pairing with T5; however, stable intermolecular-pairing interactions with Met1 or T5 (as indicated in the crystal structure¹⁰) were not identified during the production phase of these simulations (data not shown), as previously reported³⁵.

The NTH domain of Cx26 was found to be unstable (that is, rapidly unfolding) during molecular dynamics simulation in either the acetylated or non-acetylated states (Fig. 4e, Extended Data Fig. 7). The importance of this dynamical behaviour is not clear; however, we attribute this feature as a potential reason for the variation of our calculated PMFs of Cx26 compared to previous studies⁴⁷. Notably, Kwon et al. reported the Cx26 channel to be anion selective in the absence of N-terminal acetylation⁴⁷. These authors conducted an elegant set of experiments employing a grand canonical Monte-Carlo Brownian dynamics (GCMC/BD)-based approach for modelling ion conductance using a model of the Cx26 hemichannel (by extracting a single hexamer of the Cx26 intercellular channel). However, a limitation of GCMC/BD method is that the protein structure is held static, and the resulting PMFs obtained by this approach would therefore be strongly influenced by the

selected conformational state of the Cx26 NTH domain. These caveats should be considered when comparing results presented in this work.

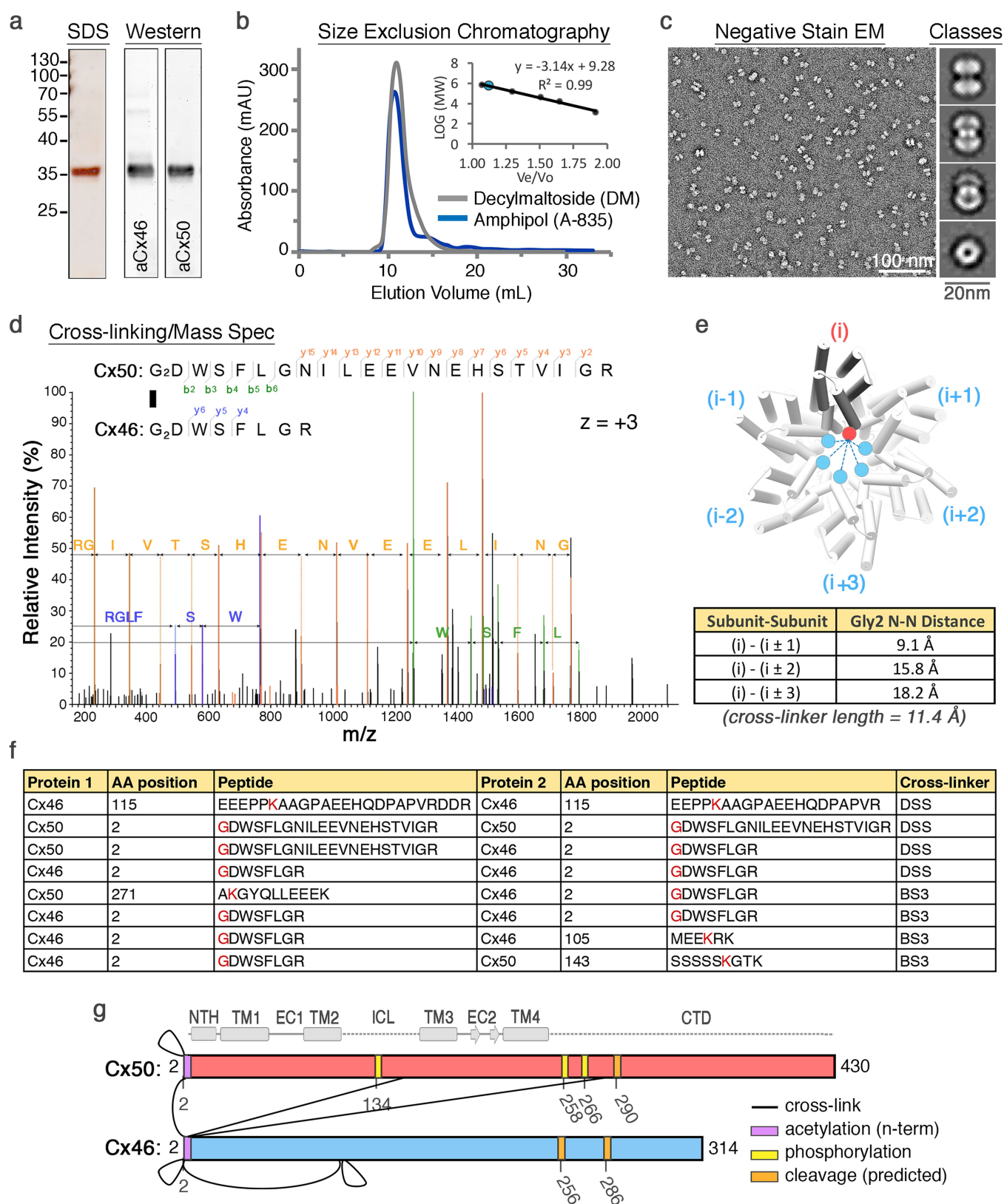
Statistical analysis. 95% confidence intervals for comparison of C_α r.m.s.f. values were calculated using a two-tailed Student's t -test. No statistical methods were used to predetermine sample size for the cryo-EM datasets. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Cryo-EM density maps have been deposited in the Electron Microscopy Data Bank under accession number EMD-9116. Coordinates for Cx46 and Cx50 atomic models have been deposited in the Protein Data Bank under accession codes 6MHQ and 6MHY. The original multi-frame micrographs have been deposited in the Electron Microscopy Public Image Archive under accession code EMPIAR-10212.

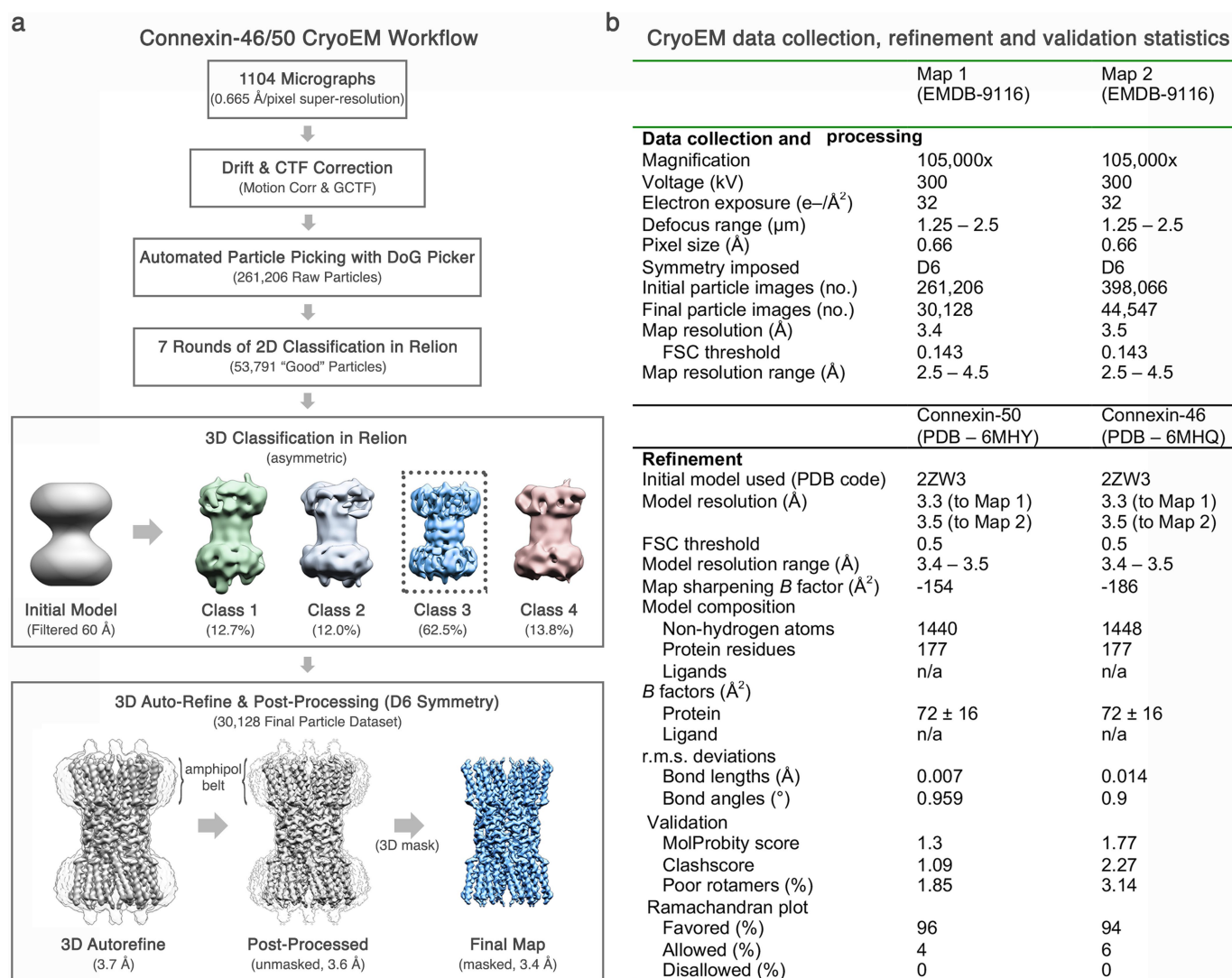
55. Kistler, J., Christie, D. & Bullivant, S. Homologies between gap junction proteins in lens, heart and liver. *Nature* **331**, 721–723 (1988).
56. Kistler, J., Schaller, J. & Sigrist, H. MP38 contains the membrane-embedded domain of the lens fiber gap junction protein MP70. *J. Biol. Chem.* **265**, 13357–13361 (1990).
57. White, T. W., Bruzzone, R., Goodenough, D. A. & Paul, D. L. Mouse Cx50, a functional member of the connexin family of gap junction proteins, is the lens fiber protein MP70. *Mol. Biol. Cell* **3**, 711–720 (1992).
58. Reichow, S. L. et al. Allosteric mechanism of water-channel gating by Ca^{2+} -calmodulin. *Nat. Struct. Mol. Biol.* **20**, 1085–1092 (2013).
59. Gold, M. G. et al. AKAP2 anchors PKA with aquaporin-0 to support ocular lens transparency. *EMBO Mol. Med.* **4**, 15–26 (2012).
60. Reichow, S. L. & Gonen, T. Noncanonical binding of calmodulin to aquaporin-0: implications for channel regulation. *Structure* **16**, 1389–1398 (2008).
61. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protocols* **1**, 2856–2860 (2006).
62. Yang, B. et al. Identification of cross-linked peptides from complex samples. *Nat. Methods* **9**, 904–906 (2012).
63. Myers, J. B. et al. The CaMKII holoenzyme structure in activation-competent conformations. *Nat. Commun.* **8**, 15742 (2017).
64. Tang, G. et al. EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
65. Mastrorade, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
66. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
67. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
68. Voss, N. R., Yoshioka, C. K., Radermacher, M., Potter, C. S. & Carragher, B. DoG Picker and TiltPicker: software tools to facilitate particle selection in single particle electron microscopy. *J. Struct. Biol.* **166**, 205–213 (2009).
69. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
70. Heymann, J. B. & Belnap, D. M. Bsoft: image processing and molecular modeling for electron microscopy. *J. Struct. Biol.* **157**, 3–18 (2007).
71. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
72. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
73. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
74. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
75. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 27–38 (1996).
76. Locke, D., Bian, S., Li, H. & Harris, A. L. Post-translational modifications of connexin26 revealed by mass spectrometry. *Biochem. J.* **424**, 385–398 (2009).
77. Grubmüller, H., Heymann, B. & Tavan, P. Ligand binding: molecular mechanics calculation of the streptavidin–biotin rupture force. *Science* **271**, 997–999 (1996).
78. Phillips, J. C. et al. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802 (2005).
79. Klauda, J. B. et al. Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *J. Phys. Chem. B* **114**, 7830–7843 (2010).
80. Best, R. B. et al. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ , ψ and side-chain χ_1 and χ_2 dihedral angles. *J. Chem. Theory Comput.* **8**, 3257–3273 (2012).
81. Zuckerman, D. M. *Statistical Physics of Biomolecules: An Introduction*, 1st edn (CRC, Boca Raton, 2010).
82. Im, W., Seefeld, S. & Roux, B. A grand canonical Monte Carlo–Brownian dynamics algorithm for simulating ion channels. *Biophys. J.* **79**, 788–801 (2000).
83. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).



Extended Data Fig. 1 | See next page for caption.

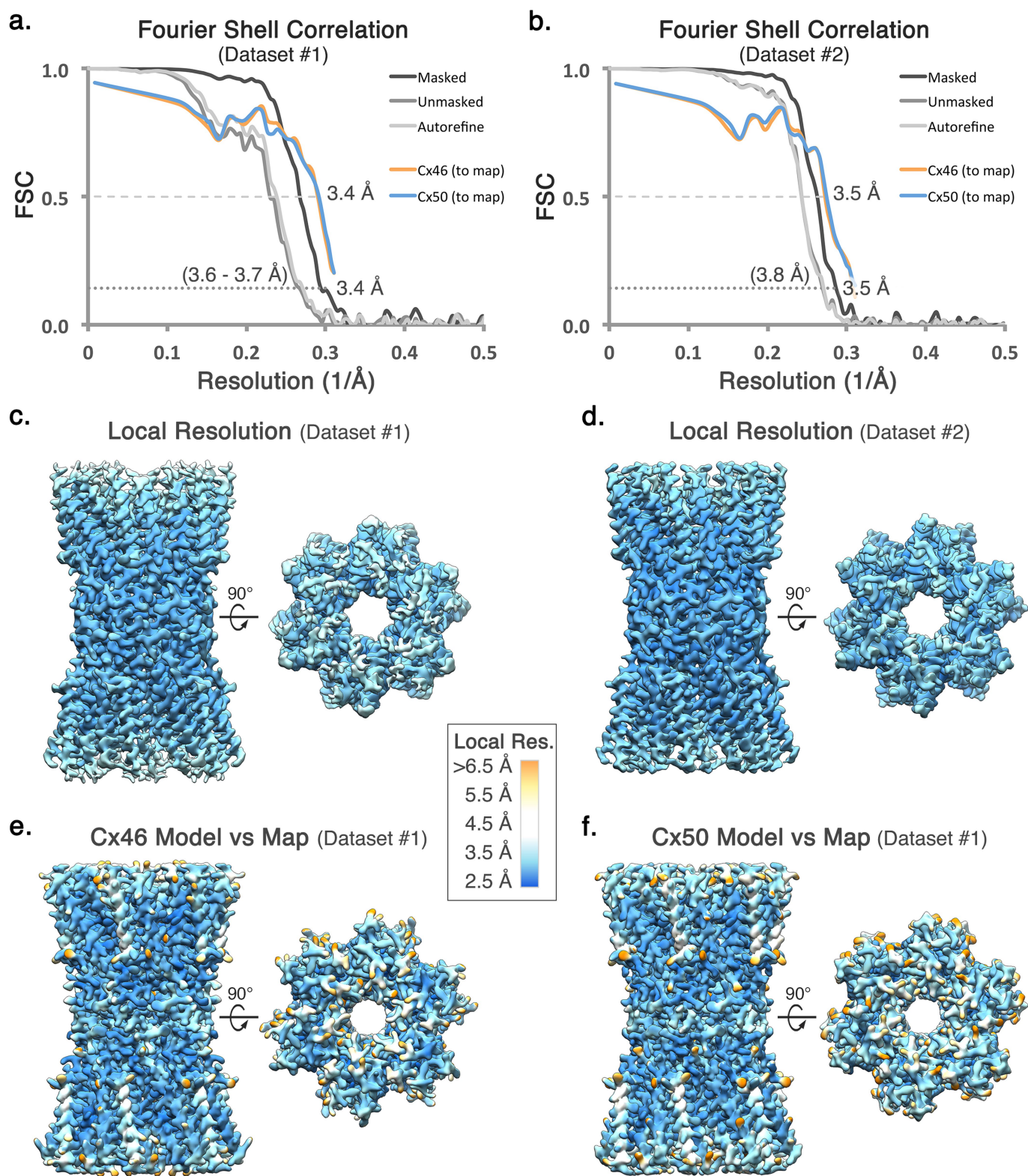
Extended Data Fig. 1 | Preliminary structural analysis of native lens Cx46/50 gap junction intercellular channels. **a**, Denaturing protein gel and western blot analysis of purified Cx46/50 (also known as MP38) isolated from lens core tissue. Protein bands corresponding to Cx46 and Cx50 co-migrate at a molecular weight of ~ 38 kDa, consistent with age-dependent proteolytic cleavage within the C-terminal domain of both isoforms¹⁶. Experiment performed 3 times with similar results. **b**, SEC elution profile of Cx46/50 gap junctions reconstituted in decyl-maltoside (DM, grey trace) or amphipol (A-835, blue trace), monitored by UV absorbance. Experiment performed more than 3 times with similar results. Inset, calibration curve ($n = 3$ runs) demonstrating that Cx46/50 elutes at an apparent molecular weight of ~ 560 kDa, consistent with the size of a dodecameric protein complex ($12 \times \sim 38$ kDa) and two micelles ($2 \times \sim 50$ kDa). **c**, Electron micrograph of negatively stained Cx46/50 gap junctions reconstituted into amphipol. Scale bar, 100 nm. Inset, representative 2D class averages of negatively stained particles (selected from 25 classes). Scale bar, 20 nm. **d–g**, Chemical cross-linking and mass spectrometry. **d**, Representative MS/MS m/z spectrum, identifying inter-subunit cross-linking at the N-terminal Gly2 positions of Cx50 and Cx46. Identified peaks in the m/z spectrum and amino acid identities are indicated (Cx50 b-ions, green; Cx50 y-ions, yellow; Cx46 y-ions, blue).

MS/MS data represent the consensus of 3 independent runs. **e**, Structural analysis of cross-linking results, showing inter-subunit distances between the symmetrically related N-terminal Gly2 positions within the connexin hemichannel, ranging from 9.1 \AA (i to $i \pm 1$), 15.8 \AA (i to $i \pm 2$) and 18.2 \AA (i to $i \pm 3$). The cross-linker spacer length is 11.4 \AA , indicating a probable (i to $i \pm 1$) arrangement of Cx50 and Cx46 within the same hemichannel, although other arrangements cannot be ruled out. **f**, Overview of identified inter-subunit cross-links between Cx50 and Cx46 assembled gap junctions. Residues in red indicate the site of primary amines involved in the cross-linking reaction using either DSS or BS3. All detected inter-subunit cross-links are between cytoplasmic domains. **g**, Schematic showing sites of inter-subunit cross-linking between Cx46 and Cx50 (black lines) and post-translational modifications identified during proteomics analysis (yellow, phosphorylation; purple, N-terminal acetylation). Met1 was determined to be removed in both Cx46 and Cx50 and the resulting N-terminal Gly2 position was identified in both acetylated and non-acetylated forms of Cx46 and Cx50, consistent with the specificity of the NatA acetylation complex²⁴. The predicted CTD cleavage sites in Cx46 and Cx50 (orange), based on previous analysis of bovine Cx46/50 isolated from lens core tissue¹⁶, are also shown. Secondary structure and domain labels are indicated for the NTH, TM1–TM4, EC1, EC2, ICL and CTD.



Extended Data Fig. 2 | Overview of cryo-EM image processing and 3D reconstruction. **a**, A total of 1,104 micrographs were collected in an automated fashion using SerialEM⁶⁵ on a 300-kV Titan Krios (dataset 1). Movie stacks were recorded using a K2 summit-direct electron detector operated in super-resolution mode and acquired with an effective pixel size of 0.665 Å. Movie stacks were corrected for drift and CTF using MotionCorr⁶⁶ and GCTF⁶⁷, respectively. An initial dataset of 261,206 raw particles was obtained using unbiased autopicking procedures in DoG Picker⁶⁸. A refined dataset of 53,791 good particles was obtained following several rounds of 2D classification and removal of 'bad' particles (or ice contamination) was done in Relion⁶⁹. 3D classification was seeded using an initial model obtained by negative-stain electron microscopy, filtered to 60 Å. A majority of particles fell into a single 3D class (~62.5% of the good particles). These 30,128 particles were used for final 3D auto-refinement

and post processing, yielding a final map at 3.4 Å resolution by gold-standard FSC (dataset 1). Dataset 2 was processed in a similar fashion from a total of 2,197 micrographs and 44,547 good particles, resulting in a final map at 3.5 Å resolution by gold-standard FSC. **b**, Summary of cryo-EM data collection, refinement and model validation statistics. Dataset 1 was used to obtain the 3.4 Å resolution reconstruction (map 1). Dataset 2 was used to obtain the 3.5 Å resolution reconstruction (map 2). Pre-processed and post-processed maps and associated masks from both datasets have been deposited in the Electron Microscopy Databank (EMD-9116). The original multi-frame micrographs have been deposited to EMPIAR (EMPIAR-10212). Coordinates for Cx50 and Cx46 atomic models have been deposited in the Protein Data Bank (6MHY and 6MHQ, respectively).

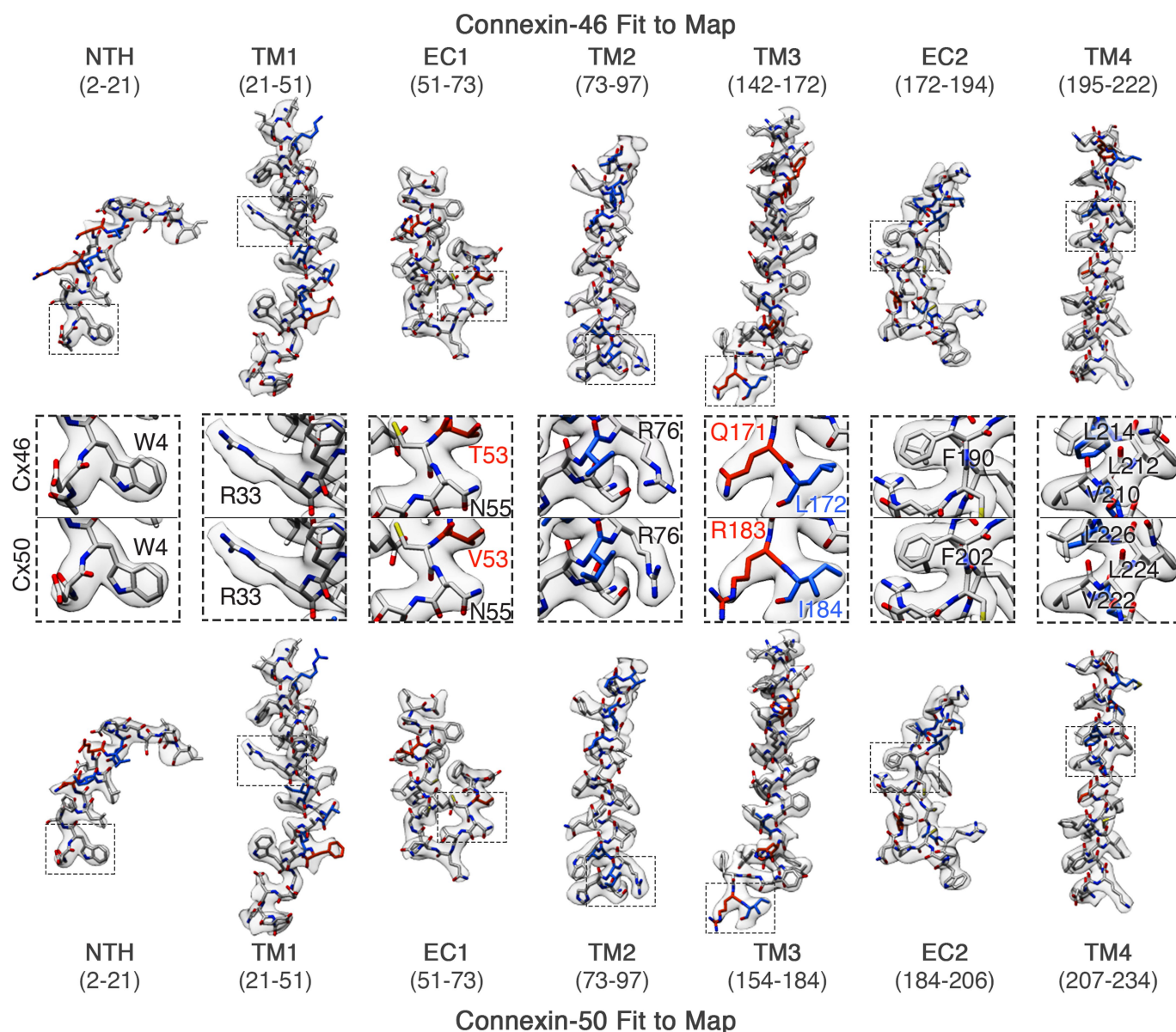


Extended Data Fig. 3 | Global and local resolution analysis. **a, b**, FSC analysis obtained from dataset 1 (**a**) and dataset 2 (**b**). Gold-standard FSC curves following auto-refinement (light grey), post-processing (grey), and masking (dark grey). The final masked maps display an overall resolution of ~ 3.4 Å (dataset 1) and ~ 3.5 Å (dataset 2), using a 0.143 cut-off. FSC curves comparing atomic models of Cx46 (orange) and Cx50 (blue) fit to the cryo-EM maps display correlation at 0.5 cut-off to a resolution of 3.4 Å (dataset 1) and 3.5 Å (dataset 2). **c, d**, Local resolution analysis using BlocRes⁷⁰, obtained for the half-maps for dataset 1 (**c**) and 2 (**d**). **e, f**, Local resolution analysis comparing the experimental density map (dataset 1)

to the calculated maps of Cx46 (**e**) and Cx50 (**f**). Local resolution ranges in **c–f** are indicated by colour (2.5–4.0 Å, blue–cyan; 4.0–5.0 Å, white; 5.0–6.5 Å, yellow–orange). Values obtained for local resolution of Cx46 and Cx50 models compared to the experimental density map are shown in Supplementary Tables 2, 3. Local resolution assessment comparing the density map to the two models indicates that the sites at which the two isoforms differ in sequence were generally less well-resolved, as compared to equivalently exposed residues at which Cx46 and Cx50 share conserved sequence.

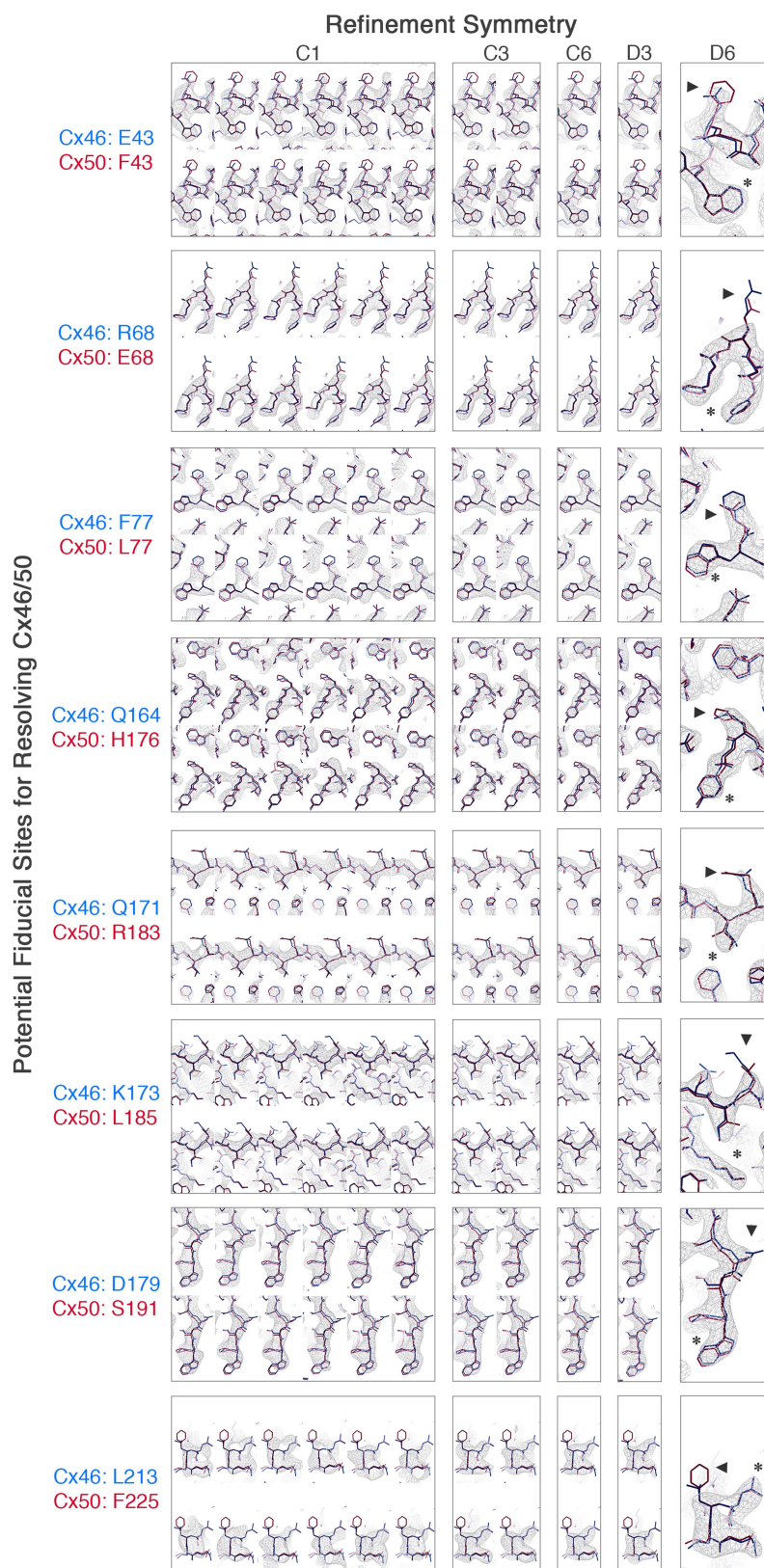
Sheep Connexin-46/50 Structural Homology

Identical (80%) = grey • Similar (8%) = blue • Different (12%) = orange



Extended Data Fig. 4 | Cx46 and Cx50 atomic models fit to the cryo-EM density maps. Segmented cryo-EM map with atomic models for sheep Cx46 and Cx50 fit to the experimental densities derived from dataset 1 (3.4 Å, D6 symmetry), including regions for TM1–TM4, EC1 and EC2. The NTH domain is fit into the map from dataset 2 (3.5 Å, D6 symmetry), which was more well-defined in this region. Cx46 (top) and Cx50 (bottom) models are coloured according to their pairwise sequence homology, as being identical (grey, 80%), similar (blue, 8%) and different (orange, 12%). Windows show magnified views corresponding to boxed regions of the segmented maps, highlighting representative side-chain densities and fit

to the atomic models. Regions of identical or similar amino acids are fit equally well by both models (for example, Cx46 L172 versus Cx50 I184, blue labels). Over regions in which the sequence of Cx46 and Cx50 differs, side-chain density is typically weaker (see also Extended Data Fig. 5). This is possibly due to the imposed averaging of two different side chains in these areas, or relative flexibility as many of these residues correspond to solvent-exposed side chains. In these areas of difference, and where electron microscopy density is present, both Cx46 and Cx50 models were typically fit equally well into the density map (for example, Cx46 T53 versus Cx50 V53 and Cx46 Q171 versus Cx50 R183, orange labels).



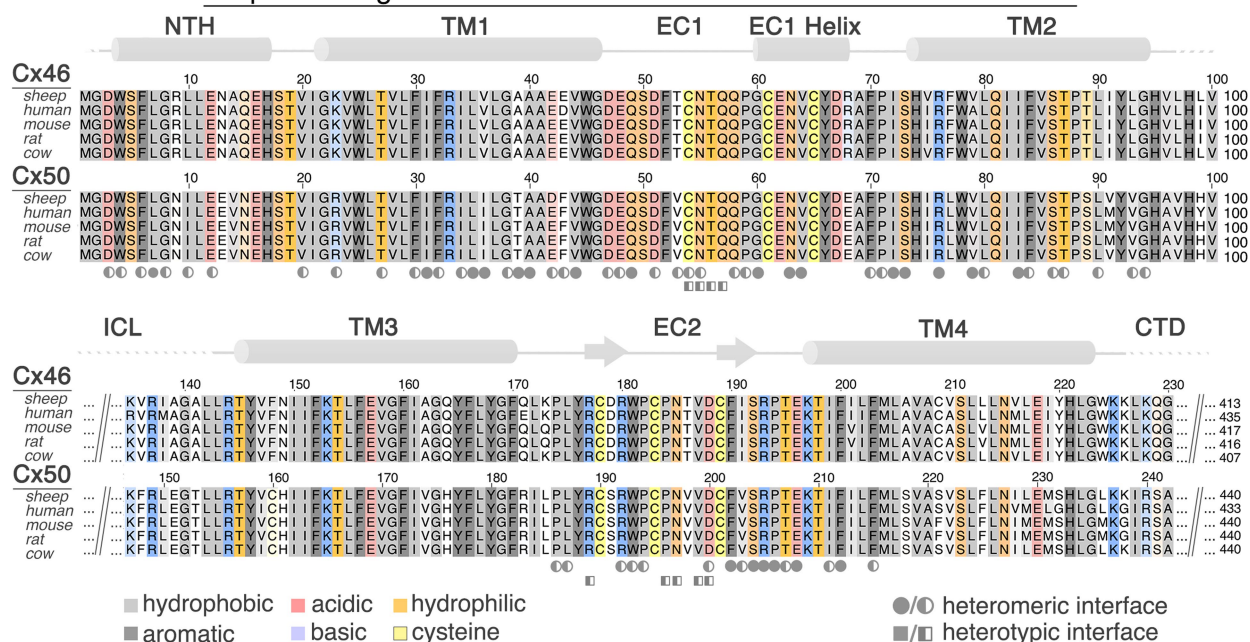
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Analysis of different symmetry refinements on the ability to resolve differences between Cx46 and Cx50. Eight sites of sequence differences involving bulky amino acids (labelled, and indicated by arrowhead) were selected as potential fiducial markers for resolving the two different isoforms following 3D refinement with various applied symmetries (C1, 4.1 Å resolution; C3, 3.9 Å resolution; C6, 3.7 Å resolution; D3, 3.7 Å resolution; D6, 3.4 Å resolution). For the applied symmetries, views are presented for each unique asymmetric subunit (boxed). Despite the modest resolution of the asymmetric (C1) reconstruction, side-chain density for bulky amino acids is typically observed at sites at which the two isoforms are conserved (asterisk). However, at the selected sites of sequence variation (arrow head) the side-

chain densities are either not well-resolved, or there was no systematic variation that indicated an ability to distinguish the two isoforms. The most resolved features at these sites of variation were obtained with D6 symmetry, and typically corresponded to regions where these different amino acids share similar structure (such as C_β positions). Although the cryo-EM density at these sites of variation were typically weak, the resolvable side-chain features throughout the rest of the map were generally enhanced when higher symmetry was applied during map refinement (indicated by asterisk), suggesting that regions of sequence similarity between Cx46 and Cx50 also share a high level of structural similarity.

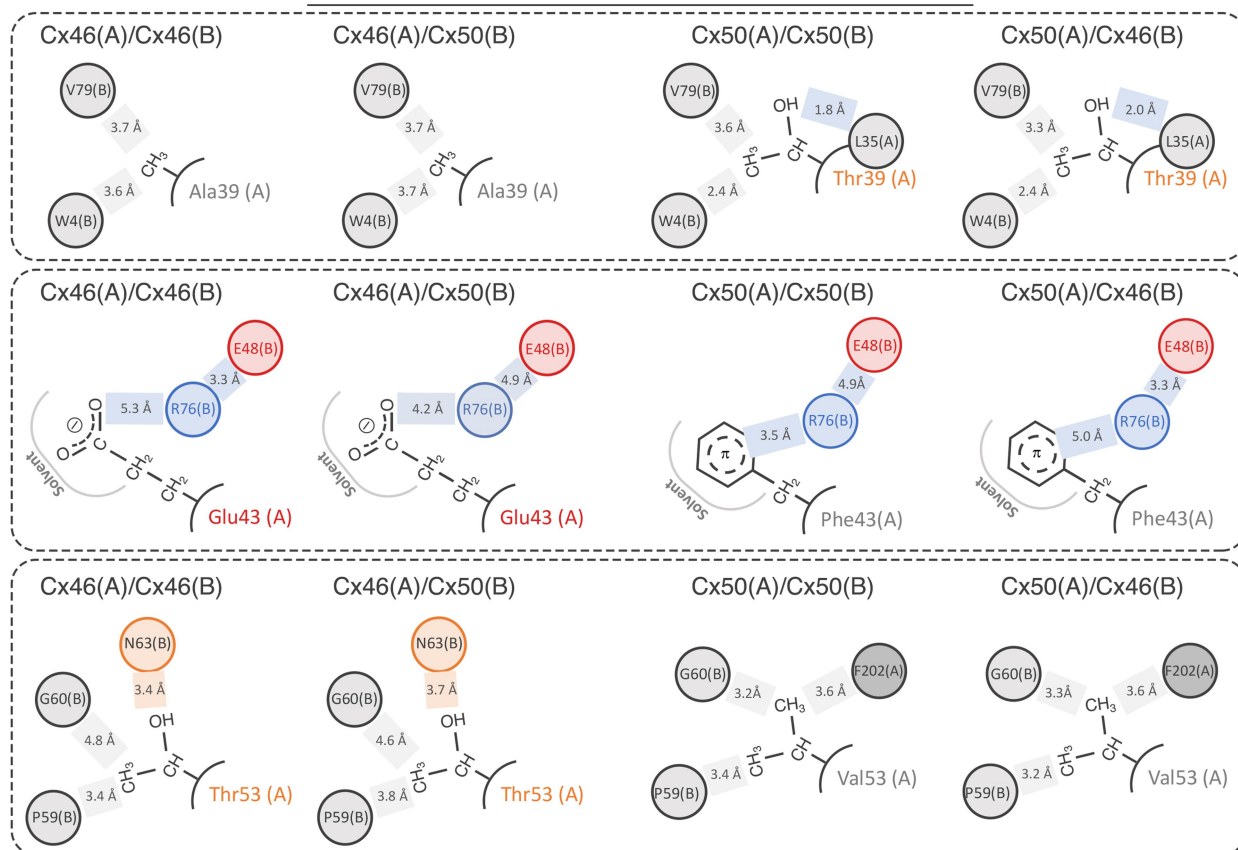
a

Sequence Alignment of Mammalian Connexin-46 and Connexin-50



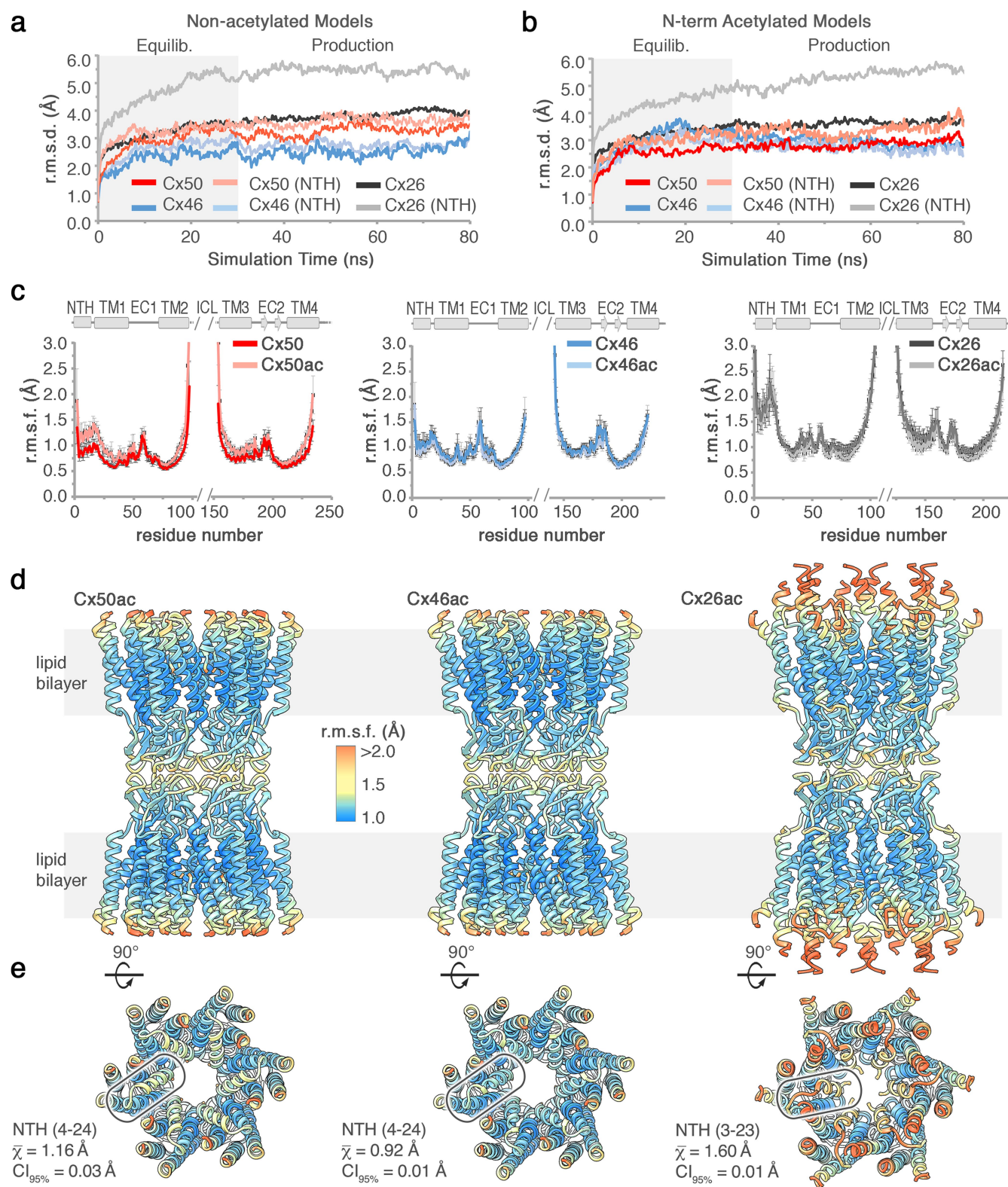
b

Variable Sites of Heteromeric Interface Interactions



Extended Data Fig. 6 | Sequence and structure conservation of Cx46 and Cx50 heteromeric/heterotypic interfaces. **a**, Multiple sequence alignment of mammalian Cx46 and Cx50 isoforms with residues contributing to heteromeric and heterotypic interfaces annotated⁸³. Circle, heteromeric interface; square, heterotypic interface; filled, $\geq 70\%$ buried; half-filled, 20–70% buried. Colouring corresponds to amino acid type (grey, hydrophobic; dark grey, aromatic; red, acidic; blue, basic; orange, hydrophilic; yellow, cysteine). Regions of sequence homology are indicated by the level of shading. Secondary structure and domain labels are indicated for the NTH domain, TM1–TM4, EC1 and EC2. Regions lacking defined structure

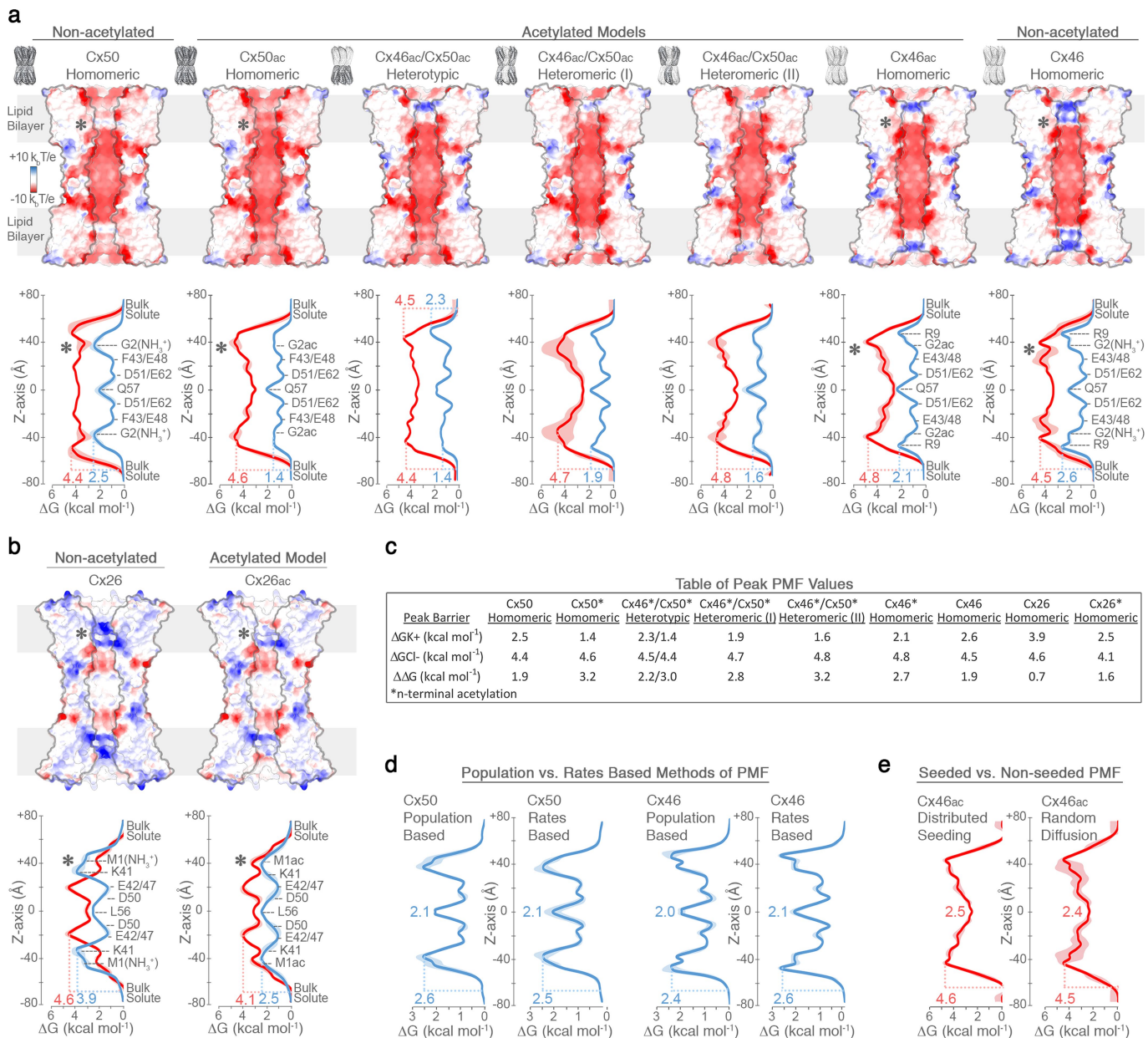
and with poor sequence homology within the intracellular loop (ICL) and C-terminal domain (CTD) have been omitted for clarity. Sheep and human Cx46 and Cx50 orthologues contain ~95% sequence identity (~98% similarity) over the structured regions of the protein. Numbering corresponds to the amino acid sequence of sheep Cx44 and Cx49 used in the main text. **b**, Illustration of homomeric and heteromeric interface interactions involving the three sites lacking conservation between Cx44 and Cx49 at this interface (positions 39, 43 and 53). Despite these sequence differences, the interactions involving these residues are generally similar (hydrophobic, grey; hydrogen bonding, orange; ion-pairing, blue).



Extended Data Fig. 7 | See next page for caption.

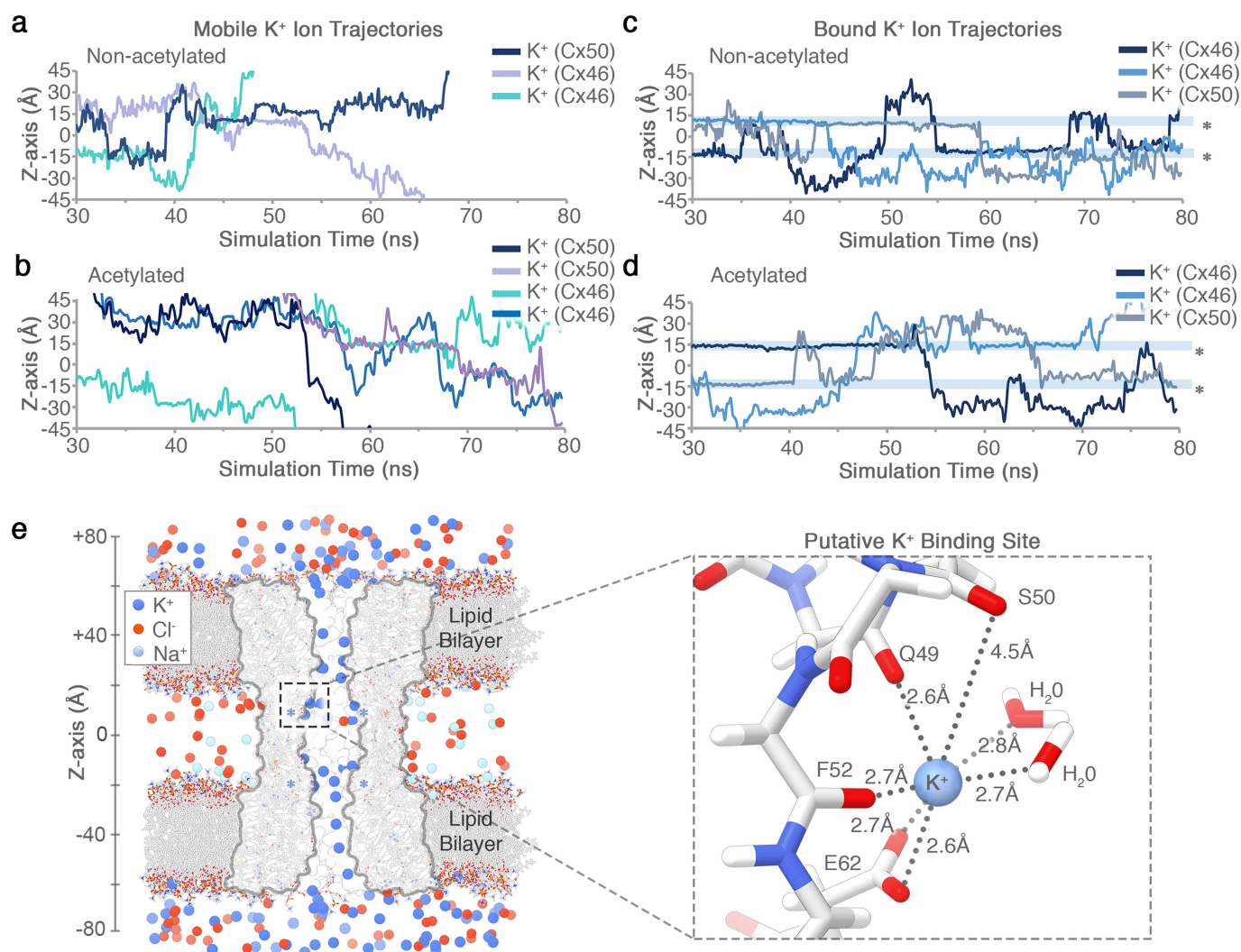
Extended Data Fig. 7 | Analysis of protein backbone dynamics during molecular dynamics equilibration and production. **a**, C_{α} r.m.s.d. analysis of equilibrium (0–30 ns) and production phases (30–80 ns) of the molecular dynamics simulations, calculated with respect to the experimental starting structure for non-acetylated models of Cx50 (red traces), Cx46 (blue traces) and Cx26 (with Met1 added; grey traces). Separate analysis for the NTH domains are shown in lighter shades. **b**, Same analysis as in **a**, for models with N-terminal acetylation added. The NTH domain of Cx26 (light grey traces) shows significantly higher r.m.s.d. values, for both non-acetylated and acetylated models. **c**, Plot of average C_{α} r.m.s.f. during the production phase of the molecular dynamics simulations for Cx50 (left, red traces), Cx46 (centre, blue traces) and Cx26 (right, grey traces). Data obtained for the N-terminal acetylated models

are shown in lighter shades. Averages are determined for the 12 subunits composing the intercellular channel. Error bars represent 95% confidence intervals ($n = 12$ subunits). Secondary structure and domain labels are indicated for the NTH, TM1–4, EC1 and EC2, and ICL (not modelled). **d**, **e**, Average r.m.s.f. values of the acetylated models mapped to the experimental starting structures of Cx50 (left), Cx46 (centre) and Cx26 (right). Colours correspond to r.m.s.f. amplitudes: 0–1.0 Å (cyan); 1.0–2.0 Å (yellow–orange), >2.0 Å (red). In **e**, a single NTH domain is circled and the average r.m.s.f. values and 95% confidence intervals (Student *t*-test) calculated over the NTH domain of each isoform are displayed ($n = 12$ subunits). The NTH domain of Cx26 shows significantly higher r.m.s.f. values, for both acetylated and non-acetylated models ($P < 0.0001$).



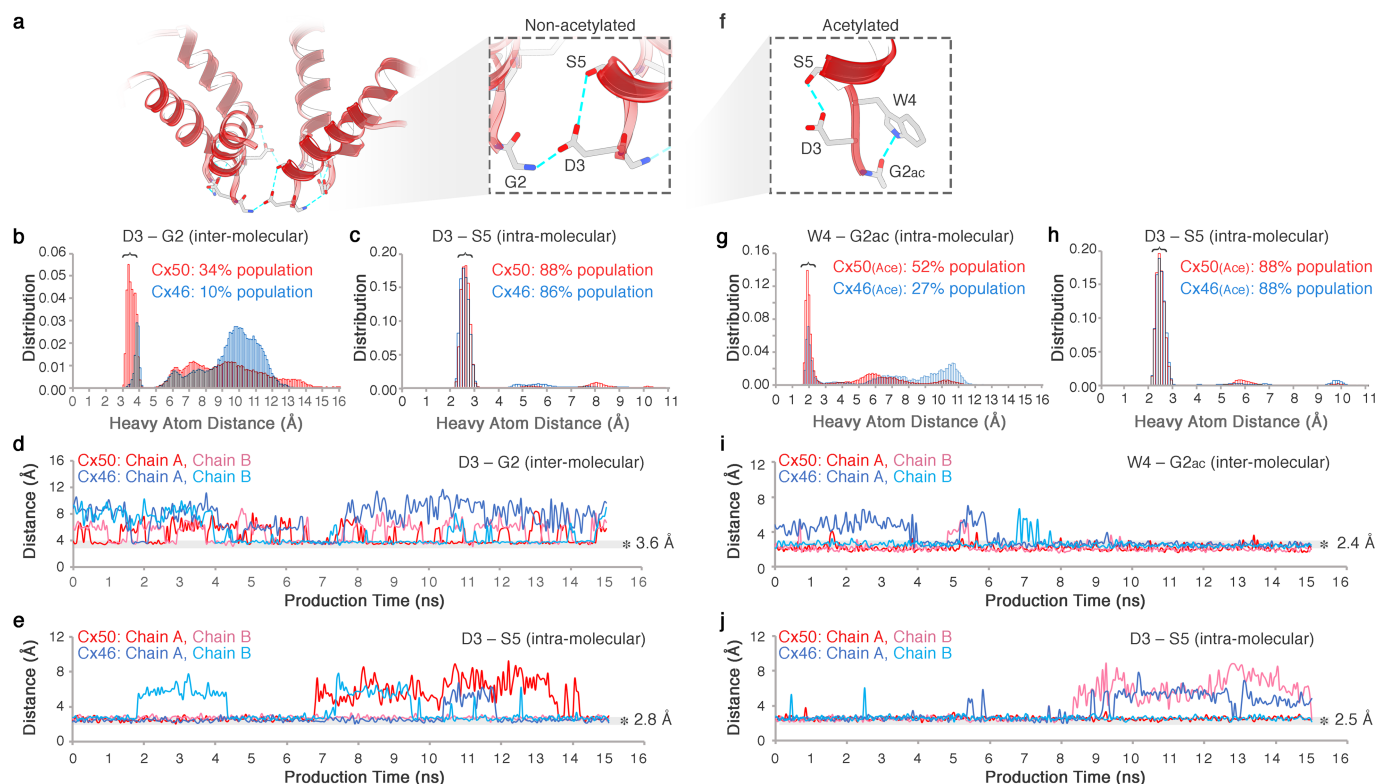
Extended Data Fig. 8 | Modulation of coulombic surface potential and K^+/Cl^- PMFs resulting from N-terminal acetylation and Cx46/50 heterotypic/heteromeric assembly. **a**, Coulombic surface potential maps (top) and PMF (bottom) obtained for a set of non-acetylated and acetylated Cx50–Cx50ac and Cx46–Cx46ac models characterized by molecular dynamics simulation. Acetylated Cx50ac and Cx46ac monomers were used to construct a heterotypic channel and two different heteromeric channels (labelled I and II). Icons at the top of each structure show the relative configurations of Cx50 (black) and Cx46 (white). The resulting coulombic surface potentials are coloured as in Fig. 3a (negative, red; neutral, white; positive, blue). Only eight subunits are shown to portray both the channel pore and subunit interfaces. An asterisk (shown in **a** and **b**) indicates the site of N-terminal acetylation, which neutralizes the positively charged N terminus. PMFs obtained for K^+ (blue traces) and Cl^- ions (red traces) are displayed directly beneath each model displayed in **a**. Free-energy maxima are labelled and pore axis (z axis) is indicated.

Traces indicate symmetrized values, with unsymmetrized values in lighter shading. In the case of the asymmetric Cx46/50 heterotypic model (middle left), PMFs represent the average from the first and last ~60 ns of simulation. **b**, Coulombic surface potential (top) and PMFs (bottom) obtained for non-acetylated Cx26 (left) and Cx26ac (right) are displayed as in **a**. **c**, Table of peak free-energy barriers for K^+ (ΔG_{K^+}) and Cl^- (ΔG_{Cl^-}) and corresponding $\Delta \Delta G$, reported as a proxy for charge selectivity. Asterisk indicates models with acetylated N terminus. **d**, **e**, Validation of methods used to construct the PMFs. **d**, Comparison of K^+ PMFs obtained for Cx50ac and Cx46ac using population states (left) or transition rates (right) (see Methods). Both methods yielded similar PMF profiles. All other PMFs were constructed using the transition rates method. **e**, Comparison of Cl^- PMFs obtained for Cx46ac using transition rates of Cl^- ions that diffused into the pore (left) and those that were randomly seeded within the pore (right). All other Cl^- PMFs were constructed using the distributed seeding approach to enhance sampling (see Methods).



Extended Data Fig. 9 | Analysis of K⁺ trajectories and putative binding site observed during molecular dynamics simulation. **a–d**, K⁺ ion trajectories obtained for Cx50 and Cx46 along the channel pore (z axis). **a, b**, Representative traces of mobile ions transiting and exiting or entering the channel pore in both acetylated (**a**) and non-acetylated (**b**) models of Cx46 and Cx50. **c, d**, K⁺ ions displaying long dwell times (~10–20 ns) localized at one or more putative binding sites within the channel pore (asterisk at $z \approx 14$ Å) in both models of Cx46 and Cx50. In **a–d**, similar results were observed from 6 independent runs using non-acetylated models (1 × 80-ns and 2 × 10-ns runs for both Cx50 and Cx46) and 13 independent runs using the acetylated models (1 × 80-ns and 6 × 10-ns runs for Cx50; and 1 × 80-ns and 7 × 10-ns runs for Cx46). **e**, Representative snapshot showing an enlarged view of the putative

K⁺ binding site identified for Cx50 and Cx46, corresponding to the region indicated by the asterisk in **c** and **d**. A single K⁺ ion is bound by a conserved set of amino acids (among Cx46/50 orthologues), coordinated by the side-chain carboxylate of Glu62 and backbone carbonyls of Gln49, Ser50 and Phe52 (identical in Cx46 and Cx50). Two transient water molecules observed coordinating the bound K⁺ ion are shown. Twelve binding sites are present within the dodecameric channel. Similar behaviour was observed from simulations using both non-acetylated and acetylated models (19 independent simulations). A functional role for this putative binding site is not yet clear, but may represent a physiologically relevant cation-binding site similar to the recently proposed Ca²⁺-binding site in Cx26¹¹.



Extended Data Fig. 10 | Dynamic hydrogen-bond network within the NTH domain observed by molecular dynamics simulation.

a–e, Analysis of hydrogen-bond interactions for non-acetylated models of Cx46 and Cx50 observed during molecular dynamics simulation. **a**, Inset, magnified view of D3 pairing with the positively charged N-terminal G2 position from a neighbouring subunit (intermolecular) and with the hydroxyl of S5 within the same subunit (intramolecular). The D3–G2 interactions are dynamically formed and broken during molecular dynamics simulation, whereas the intramolecular D3–S5 hydrogen bond is relatively stable (as shown in **b–e**). **b, c**, Population statistics of inter-atomic distances involving D3 (C_{γ}) and G (N) of the neighbouring chain (**b**) and Ser5 (H_{γ}) of the same chain (**c**), extracted from molecular dynamics simulation production runs of Cx46 (blue histogram) and Cx50 (red histogram). For D3 and G2, heavy atoms were chosen as proxies to monitor hydrogen-bonding interactions involving equivalent rotameric donor-acceptor configurations. The population centred at ~3.6 Å (**b**) and ~2.8 Å (**c**) are considered to be within hydrogen-bond distance. **d, e**, Trajectories extracted from molecular dynamics simulation of Cx46 (blue traces) and Cx50 (red traces) showing the dynamical behaviour of the D3–G2 intermolecular charge pairing (**d**) and D3–S5 intramolecular hydrogen bonding (**e**). The dwell times showing hydrogen-bond pairing (~3.6 Å in **d**; and ~2.8 Å in **e**) are indicated with transparent grey shading. In the Cx26 crystal structure, the equivalent D2 site is modelled in hydrogen-bond distance to a neighbouring T5 site (Cx26 numbering)¹⁰,

but this intermolecular interaction is rapidly broken during molecular dynamics simulations and does not appear to reform within the timescale of our molecular dynamics experiments, and instead forms a stable intramolecular interaction with T5, as previously described³⁵ (data not shown). **f–j**, Analysis of hydrogen-bond interactions observed during molecular dynamics simulation for Cx46ac and Cx50ac modelled with the N-terminal G2 position acetylated. **f**, Inset, magnified view of acetylated G2ac position hydrogen bonded to the indole ring of W4 from the same subunit (intramolecular) and the same intramolecular D3–S5 hydrogen-bond interaction observed in the non-acetylated channel. **g, h**, Population statistics of inter-atomic distances involving W4 (N_{ϵ}) and G2ac (acetyl carbonyl) (**g**) and D3 (C_{γ}) distance to Ser5 (H_{γ}) of the same chain (**h**), extracted from molecular dynamics simulation production runs of Cx46ac (blue histogram) and Cx50ac (red histogram). **i, j**, Trajectories extracted from molecular dynamics simulation of Cx46ac (blue traces) and Cx50ac (red traces) showing the dynamical behaviour of the W4–G2ac hydrogen-bond pairing (**i**) and D3–S5 intramolecular hydrogen bonding (**j**). The dwell times showing hydrogen-bond pairing (~2.4 Å in **i**, and ~2.5 Å in **j**) are indicated with transparent grey shading. For clarity, only the first 15 ns of the production period is shown (**d, e, i, j**). Similar results were observed from 6 independent runs using non-acetylated models (1 × 80-ns and 2 × 10-ns runs for both Cx50 and Cx46) and 13 independent runs using the acetylated models (1 × 80-ns and 6 × 10-ns runs for Cx50; and 1 × 80-ns and 7 × 10-ns runs for Cx46).

Extreme ^{13}C , ^{15}N and ^{17}O isotopic enrichment in the young planetary nebula K4-47

D. R. Schmidt¹, N. J. Woolf¹, T. J. Zega² & L. M. Ziurys^{1,3,4*}

Carbon, nitrogen and oxygen are the three most abundant elements in the Galaxy after hydrogen and helium. Whereas hydrogen and helium were created in the Big Bang, carbon, nitrogen and oxygen arise from nucleosynthesis in stars. Of particular interest^{1,2} are the isotopic ratios $^{12}\text{C}/^{13}\text{C}$, $^{14}\text{N}/^{15}\text{N}$ and $^{16}\text{O}/^{17}\text{O}$ because they are effective tracers of nucleosynthesis and help to benchmark the chemical processes that occurred in primitive interstellar material as it evolved into our Solar System³. However, the origins of the rare isotopes ^{15}N and ^{17}O remain uncertain, although novae and very massive stars that explode as supernovae are postulated^{4–6} to be the main sources of ^{15}N . Here we report millimetre-wavelength observations of the young bipolar planetary nebula K4-47 that indicate another possible source for these isotopes. We identify various carbon-bearing molecules in K4-47 that show that this object is carbon-rich, and find unusually high enrichment in rare carbon (^{13}C), oxygen (^{17}O) and nitrogen (^{15}N) isotopes: $^{12}\text{C}/^{13}\text{C} = 2.2 \pm 0.8$, $^{16}\text{O}/^{17}\text{O} = 21.4 \pm 10.3$ and $^{14}\text{N}/^{15}\text{N} = 13.6 \pm 6.5$ (uncertainties are three standard deviations); for comparison, the corresponding solar ratios⁷ are 89.4 ± 0.2 , $2,632 \pm 7$ and 435 ± 57 . One possible interpretation of these results is that K4-47 arose from a *J*-type asymptotic giant branch star that underwent a helium-shell flash (an explosive nucleosynthetic event that converts large quantities of helium to carbon and other elements), enriching the resulting planetary nebula in ^{15}N and ^{17}O and creating its bipolar geometry. Other possible explanations are that K4-47 is a binary system or that it resulted from a white dwarf merger, as has been suggested for object CK Vul⁸. These results suggest that nucleosynthesis of carbon, nitrogen and oxygen is not well understood and that the classification of certain stardust grains must be reconsidered.

The planetary nebula phase, which follows the asymptotic giant branch (AGB) track, marks the end of the stellar life cycle as the star ejects most of its mass and evolves into a white dwarf, a strong emitter of ultraviolet radiation. This radiation subsequently ionizes the ejected stellar material, creating a bright nebula. K4-47, a young planetary nebula with an age of 400–900 years⁹, is particularly interesting because of its kinematic structure. K4-47 has a highly collimated bipolar outflow, visible in the emission of vibrationally excited H_2 , as well as a hot central region traced by highly excited atomic lines such as $[\text{O III}]$, as shown in Fig. 1^{9,10}. Furthermore, CO, HCO^+ and CS have been observed at millimetre wavelengths in K4-47 by our group¹¹.

We conducted millimetre-wavelength observations of K4-47 to further probe its molecular content using the new Atacama Large Millimeter/submillimeter Array (ALMA) prototype 12-m antenna and the Submillimetre Telescope (SMT) of the Arizona Radio Observatory (ARO) between April 2016 and June 2018, as well as the 30-m telescope of the Institut de Radioastronomie Millimétrique (IRAM) between 16 and 18 December 2017. The ARO and IRAM data were reduced using UNIPOPS and CLASS, respectively. In both cases, individual 6-min scans were co-added to improve the signal-to-noise ratio. First- to third-order baselines were subtracted from the spectra, which were typically measured with a resolution of 1 MHz (see Extended Data

Table 1 for the spectral line strengths). A non-LTE (local thermodynamic equilibrium) radiative transfer analysis of the observed spectral lines was performed using the RADEX code¹², which accounts for any optical-depth effects. The code predicts observed line intensities to determine column densities for each molecule—a measure of their abundance, given known physical conditions (temperature and gas densities; see Methods)^{11,13–15}. The distribution of molecular material, and therefore any source coupling factor, is well constrained by previous H_2 mapping¹⁰.

Our millimetre-wavelength observations showed that K4-47 is rich in gas-phase molecules, including HCN, HNC, CCH, CN and HC_3N ^{11,13–15}, as well as the isotopically rare molecules H^{13}CN , HN^{13}C , ^{13}CS , ^{13}CN , HC^{15}N , H^{13}CCCN , HC^{13}CCN , HCC^{13}CN and C^{17}O , as shown in Fig. 2. These molecules were detected via multiple millimetre-wavelength rotational transitions, typically observed at wavelengths of 3 mm and 1 mm.

The presence of such molecules in K4-47 is unexpected. The molecular content of planetary nebulae has always been considered to be extremely low¹⁶ because the ultraviolet radiation from the emerging white dwarf is thought to photodissociate most molecular species that are prominent in the previous AGB phase. Furthermore, the detection of these chemical compounds indicates a carbon-rich environment, reflecting that $\text{C} > \text{O}$ in the progenitor star¹⁷. Comparison of molecular column densities for HCN, HNC, CN, HC_3N , CO and CS and their corresponding isotopologues results in striking C, N and O isotope ratios: $^{14}\text{N}/^{15}\text{N} = 13.6 \pm 6.5$, $^{16}\text{O}/^{17}\text{O} = 21.4 \pm 10.3$ and $^{12}\text{C}/^{13}\text{C} = 2.2 \pm 0.8$

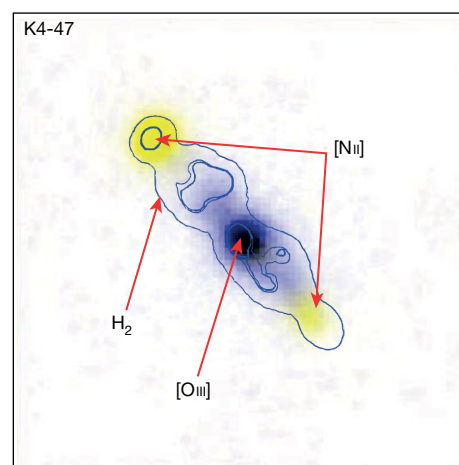


Fig. 1 | Composite image of the young planetary nebula K4-47. The contours of the H_2 vibrational transition $v = 1 \rightarrow 0 \text{ S}(1)^{10}$, which traces molecular material, are shown, as well as $[\text{O III}]$ and $[\text{N II}]$ emission, which highlight the highly ionized gas near the central star and the low-ionization edges of the bipolar flow, respectively⁹. The H_2 data are from Fig. 2 of ref. ¹⁰ and the $[\text{N II}]$ and $[\text{O III}]$ data from figure 6 of ref. ⁹. The composite image was generated using ImageJ.

¹Department of Astronomy, Steward Observatory, University of Arizona, Tucson, AZ, USA. ²Department of Planetary Science, Lunar and Planetary Laboratory, University of Arizona, Tucson, AZ, USA. ³Department of Chemistry and Biochemistry, University of Arizona, Tucson, AZ, USA. ⁴Arizona Radio Observatory, Steward Observatory, University of Arizona, Tucson, AZ, USA.

*e-mail: lziurys@email.arizona.edu

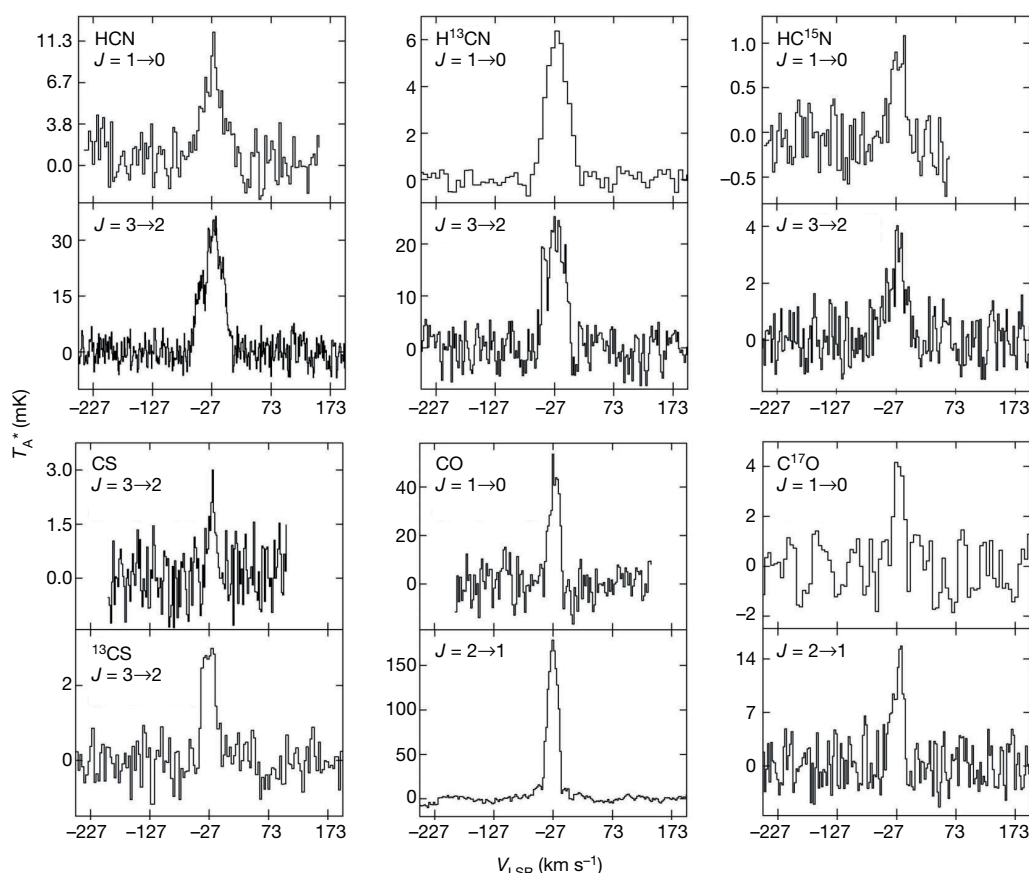


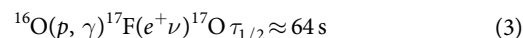
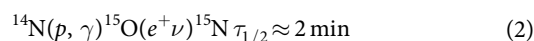
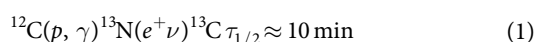
Fig. 2 | Spectra showing the isotopic enrichment in planetary nebula K4-47. The molecule and the total angular momentum quantum number J corresponding to each rotational transition are displayed in each panel. The spectra are plotted in terms of intensity (T_A^*), given here in millikelvin, versus velocity with respect to the local standard of rest (V_{LSR}), in

kilometres per second. The $J = 1 \rightarrow 0$ transitions of all molecules were observed with the new ARO ALMA 12-m prototype antenna, whereas the $J = 3 \rightarrow 2$ lines were measured with the ARO SMT or the IRAM 30-m. The spectral resolution is 1–2 MHz.

(3σ uncertainties; see Extended Data Table 1). Chemical fractionation cannot explain these ratios, because the gas temperatures in K4-47 are too high (≥ 55 K). Although $^{12}\text{C}/^{13}\text{C}$ ratios as low as about 2–4 have been found in envelopes of red giants and J -type stars^{18,19}, to our knowledge, the observed $^{14}\text{N}/^{15}\text{N}$ and $^{16}\text{O}/^{17}\text{O}$ ratios are the lowest found thus far in interstellar gas. Our observations of K4-47 suggest that there may be a new, undiscovered source of ^{15}N and ^{17}O .

The high carbon abundance, combined with the extremely low isotope ratios, indicates that K4-47 may have had a J -type progenitor star that underwent an explosive He-shell flash. J -type stars are a subclass of carbon-rich AGB stars, representing about 15% of the population²⁰; they have low $^{12}\text{C}/^{13}\text{C}$ ratios close to the CNO-cycle equilibrium values^{18,19} (about 3–19; see Table 1). Most carbon-rich AGB stars have much higher $^{12}\text{C}/^{13}\text{C}$ ratios, typically²¹ >40 , because helium-shell burning enhances ^{12}C relative to ^{13}C . J -type stars also exhibit lower-than-solar $^{16}\text{O}/^{17}\text{O}$ ratios (about 300–900)^{18,19} and, in some cases, reduced $^{14}\text{N}/^{15}\text{N}$ ratios of 153 ± 50 to 282 ± 40 , as found for Y CVn and RY Dra, respectively (Table 1). It has been previously proposed²² that AGB stars can generate some ^{15}N .

^{13}C , ^{15}N and ^{17}O can be produced by adding protons to ^{12}C , ^{14}N and ^{16}O , respectively, at very high temperatures ($T \approx 10^8$ K), which creates the unstable isotopes ^{13}N , ^{15}O and ^{17}F , correspondingly²³. These nuclei then decay relatively quickly (half-lives $\tau_{1/2} < 10$ min) to ^{13}C , ^{15}N and ^{17}O , respectively, by emitting a β particle (e^+)²³:



The intermediate nuclei must be quickly removed from the region of proton addition to mitigate other ‘hot CNO’ reactions, such as $^{13}\text{N}(p, \gamma)^{14}\text{O}$, which decays²³ to ^{14}N . All of these processes could occur in an energetic helium-shell flash in a J -type star. In the simple shell-flash scenario, helium fuses into ^{12}C , which is subsequently injected into the hydrogen-burning shell. The hot carbon then reacts to form ^{13}C via (1) and elevates the temperatures, initiating (2) and (3). The β decay reactions leading to ^{13}C , ^{15}N and ^{17}O occur after the material has been expelled from the H-shell, before competing reactions can occur. Such an explosive process could give rise to the bipolar flows found in many planetary nebulae, which are currently a mystery. If the ejection is less energetic, the production of ^{15}N and ^{17}O will not be as high as in K4-47, and the material may simply mix into the stellar envelope, creating more ordinary J -stars, such as Y CVn, that have some enhancement of ^{15}N . Accurate hydrodynamic stellar modelling is certainly needed to examine this highly qualitative picture.

On the other hand, it is thought that binary stars may be necessary to produce bipolar planetary nebulae²⁴. Current theoretical models appear to have difficulty in producing bipolar outflows in planetary nebulae from single stars, whereas such outflows can readily be formed in a binary system²⁴. In the ‘common envelope’ scheme, one star in a binary system could undergo thermal pulses and become carbon-rich before the merger. Therefore, an alternative explanation is that K4-47 results from a binary star interaction, possibly a merger with a white dwarf. The probability of binary mergers leading to planetary nebulae is not known, as many such objects have not been studied in similar detail to K4-47.

Table 1 | $^{12}\text{C}/^{13}\text{C}$, $^{14}\text{N}/^{15}\text{N}$ and $^{16}\text{O}/^{17}\text{O}$ ratios in K4-47 and related sources^a

| Ratio | Molecular tracers | K4-47 | CK Vul ^b | J-type stars | Nova grains | Solar |
|-------------------------------|-------------------------------------|-----------------|---------------------|------------------------|-----------------------|------------------|
| $^{12}\text{C}/^{13}\text{C}$ | HCN, CS, CN, HNC, HC ₃ N | 2.2 ± 0.8 | 3.8 ± 1.0 | 3–19 ^c | 4–9 ^d | 89.4 ± 0.2^e |
| $^{14}\text{N}/^{15}\text{N}$ | HCN | 13.6 ± 6.5 | 20.0 ± 10.0 | 153–282 | 5–20 ^d | 435 ± 57^e |
| $^{16}\text{O}/^{17}\text{O}$ | CO | 21.4 ± 10.3 | >110 | 270–850 ^{c,f} | 23–2,591 ^g | $2,632 \pm 7^e$ |

^aUncertainties are 3σ ; molecular tracers apply to K4-47 only.^bFrom Kamiński et al.⁸.^cFrom Abia et al.¹⁸.^dFrom Lodders & Amari²⁸.^eFrom Asplund et al.⁷.^fFrom Harris et al.¹⁹.^gFrom Iliadis et al.²⁹.

The only other astronomical object that resembles K4-47 is the enigmatic CK Vul⁸, which is currently characterized as a white dwarf merger^{8,25}. This carbon-rich object has a bipolar outflow and a central ionizing source with $T > 50,000$ K, as well as a similar set of molecules, indicating $^{12}\text{C}/^{13}\text{C} = 3.8 \pm 1.0$, $^{14}\text{N}/^{15}\text{N} = 20 \pm 10$ and $^{16}\text{O}/^{17}\text{O} > 110$. The molecular content of both objects is at least $0.5 M_{\odot}$ (M_{\odot} , mass of the Sun)²⁶. The outflow velocities of about 400 km s^{-1} in CK Vul, as traced by CO and HCN, however, are considerably higher than those in K4-47 (about $60\text{--}80 \text{ km s}^{-1}$). Such velocities are too low to classify K4-47 as a nova shell (velocities of about $370\text{--}850 \text{ km s}^{-1}$)²⁷.

Aside from K4-47 and CK Vul, similarly low $^{12}\text{C}/^{13}\text{C}$, $^{14}\text{N}/^{15}\text{N}$ and $^{16}\text{O}/^{17}\text{O}$ ratios have been found in presolar grains—small, $0.1\text{--}20\text{-}\mu\text{m}$ -sized particles extracted from meteorites²⁸. These grains are known to predate the Solar System and originate in the circumstellar envelopes of stars that have long since died. Presolar grains composed of silicon carbide (SiC) have been assigned as originating in AGB stars (including J-type) or novae/supernovae, on the basis of their $^{12}\text{C}/^{13}\text{C}$ and $^{14}\text{N}/^{15}\text{N}$ ratios, as shown in Fig. 3²⁸. Some SiC grains exhibit extremely low ratios²⁸ of $^{12}\text{C}/^{13}\text{C} \approx 4\text{--}9$ and $^{14}\text{N}/^{15}\text{N} \approx 5\text{--}20$, comparable to those of K4-47 and CK Vul. Furthermore, presolar oxide/silicate grains have been found²⁹ to have $^{16}\text{O}/^{17}\text{O}$ ratios as low as about 20, with solar $^{18}\text{O}/^{16}\text{O}$ (we note that SiC grains contain too little oxygen for determination of the $^{16}\text{O}/^{17}\text{O}$ ratio). On the basis of nucleosynthesis models, both types of grains have been tentatively assigned to novae³⁰. However, the isotopic compositions of these grains do not

actually match predictions of nova models^{29,30}. This discrepancy has led to complicated scenarios involving mixing of solar material into nova ejecta to produce the observed ratios³⁰. Moreover, on the basis of silicon isotopes, the grains must come from less-common novae with white dwarf stars rich in oxygen and neon (ONe novae), which do not effectively form carbon-rich, SiC-type dust grains. Supernovae have been proposed as an alternative source, but models cannot explain the low $^{16}\text{O}/^{17}\text{O}$ ratios found in the grains⁵.

The $^{14}\text{N}/^{15}\text{N}$ and $^{12}\text{C}/^{13}\text{C}$ ratios that we observe for K4-47 and CK Vul place these objects clearly among the ‘putative nova’ SiC presolar grains, as shown in Fig. 3. Furthermore, the $^{16}\text{O}/^{17}\text{O}$ ratio in K4-47 matches well those measured in oxide/silicate presolar grains also attributed to novae³⁰. These remarkable results suggest that the nova presolar grains actually arise from stars other than novae, perhaps explosive J-type stars that produce large amounts of ^{15}N and ^{17}O , which evolve into planetary nebulae, as represented by K4-47. It is noteworthy that presolar A+B-type SiC grains are thought to come from J-type stars (Fig. 3). The A+B grain population appears directly above that considered to arise from novae. Thus, the putative nova grains may be simply an extension of A+B grains—all created from J-type stars.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0763-1>.

Received: 1 August 2018; Accepted: 31 October 2018;

Published online 19 December 2018.

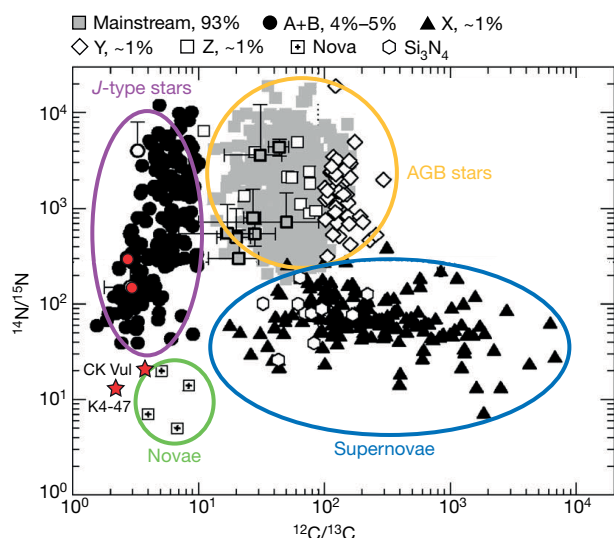


Fig. 3 | $^{14}\text{N}/^{15}\text{N}$ versus $^{12}\text{C}/^{13}\text{C}$ in SiC grains, K4-47 and CK Vul. The presolar SiC grains are identified as arising from J-type stars for the A+B-type grains (circled in purple), AGB stars for the mainstream, Y and Z types (circled in yellow), supernovae for the X and Si₃N₄ types (circled in blue) and putative nova grains (circled in green); figure adapted from figure 6 of ref. ²⁸. Data for Y CVn and RY Dra are shown with red circles and for K4-47 and CK Vul with red stars. The plot suggests that at least some nova grains arise from extreme J-type stars. Stellar data are plotted with error bars that indicate the 3σ confidence level.

- Wilson, T. L. & Rood, R. T. Abundances in the interstellar medium. *Annu. Rev. Astron. Astrophys.* **32**, 191–226 (1994).
- Adande, G. R. & Ziurys, L. M. Millimeter-wave observations of CN and HNC and their ^{15}N isotopologues: a new evaluation of the $^{14}\text{N}/^{15}\text{N}$ ratio across the Galaxy. *Astrophys. J.* **744**, 194 (2012).
- Busemann, H. et al. Interstellar chemistry recorded in organic matter from primitive meteorites. *Science* **312**, 727–730 (2006).
- José, J. & Hernanz, M. Nucleosynthesis in classical novae: CO versus ONe white dwarfs. *Astrophys. J.* **494**, 680–690 (1998).
- Pignatari, M. et al. Carbon-rich presolar grains from massive stars: subsolar $^{12}\text{C}/^{13}\text{C}$ and $^{14}\text{N}/^{15}\text{N}$ ratios and the mystery of ^{15}N . *Astrophys. J.* **808**, L43 (2015).
- Romano, D., Matteucci, F., Zhang, Z. Y., Papadopoulos, P. P. & Ivison, R. J. The evolution of CNO isotopes: a new window on cosmic star formation history and the stellar IMF in the age of ALMA. *Mon. Not. R. Astron. Soc.* **470**, 401–415 (2017).
- Asplund, M., Grevesse, N., Sauval, A. J. & Scott, P. The chemical composition of the Sun. *Annu. Rev. Astron. Astrophys.* **47**, 481–522 (2009).
- Kamiński, T. et al. Organic molecules, ions, and rare isotopologues in the remnant of the stellar-merger candidate, CK Vulpeculae (Nova 1670). *Astron. Astrophys.* **607**, A78 (2017).
- Corradi, R. et al. High-velocity collimated outflows in planetary nebulae: NGC 6337, He 2–186, and K4–47. *Astrophys. J.* **535**, 823–832 (2000).
- Akras, S., Gonçalves, D. R. & Ramos-Larios, G. H₂ in low-ionization structures of planetary nebulae. *Mon. Not. R. Astron. Soc.* **465**, 1289–1296 (2017).
- Edwards, J. L., Cox, E. G. & Ziurys, L. M. Millimeter observations of CS, HCO⁺, and CO toward five planetary nebulae: following molecular abundances with nebular age. *Astrophys. J.* **791**, 79 (2014).
- van der Tak, F. F. S., Black, J. H., Schöier, F. L., Jansen, D. J. & van Dishoeck, E. F. A computer program for fast non-LTE analysis of interstellar line spectra. *Astron. Astrophys.* **468**, 627–635 (2007).
- Schmidt, D. R. & Ziurys, L. M. Hidden molecules in planetary nebulae: new detections of HCN and HCO⁺ from a multi-object survey. *Astrophys. J.* **817**, 175 (2016).

14. Schmidt, D. R. & Ziurys, L. M. New detections of HNC in planetary nebulae: evolution of the [HCN]/[HNC] ratio. *Astrophys. J.* **835**, 79 (2017).
15. Schmidt, D. R. & Ziurys, L. M. New identifications of the CCH radical in planetary nebulae: a connection to C₆₀? *Astrophys. J.* **850**, 123 (2017).
16. Redman, M. P., Viti, S., Cau, P. & Williams, D. A. Chemistry and clumpiness in planetary nebulae. *Mon. Not. R. Astron. Soc.* **345**, 1291–1296 (2003).
17. Edwards, J. L. & Ziurys, L. M. Sulfur- and silicon-bearing molecules in planetary nebulae: the case of M2–48. *Astrophys. J.* **794**, L27 (2014).
18. Abia, C., Hedrosa, R. P., Domínguez, I. & Straniero, O. The puzzle of the CNO isotope ratios in asymptotic giant branch carbon stars. *Astron. Astrophys.* **599**, A39 (2017).
19. Harris, M. J., Lambert, D. L., Hinkle, K. H., Gustafsson, B. & Eriksson, K. Oxygen isotopic abundances in evolved stars. III. 26 carbon stars. *Astrophys. J.* **316**, 294–304 (1987).
20. Abia, C. & Isern, J. The chemical composition of carbon stars. II. The J-type stars. *Astrophys. J.* **536**, 438–449 (2000).
21. Milam, S. N., Woolf, N. J. & Ziurys, L. M. Circumstellar ¹²C/¹³C isotope ratios from millimeter observations of CN and CO: mixing in carbon- and oxygen-rich stars. *Astrophys. J.* **690**, 837–849 (2009).
22. Hedrosa, R. P. et al. Nitrogen isotopes in asymptotic giant branch carbon stars and presolar SiC grains: a challenge for stellar nucleosynthesis. *Astrophys. J.* **768**, L11 (2013).
23. Wiescher, M., Görres, J., Uberseder, E., Imbriani, G. & Pignatari, M. The cold and hot CNO cycles. *Annu. Rev. Nucl. Part. Sci.* **60**, 381–404 (2010).
24. De Marco, O. The origin and shaping of planetary nebulae: putting the binary hypothesis to the test. *Publ. Astron. Soc. Pacif.* **121**, 316–342 (2009).
25. Kamiński, T. et al. Astronomical detection of radioactive molecule ²⁶AlF in the remnant of an ancient explosion. *Nat. Astron.* **2**, 778–783 (2018).
26. Kamiński, T. et al. Nuclear ashes and outflow in the eruptive star Nova Vul 1670. *Nature* **520**, 322–324 (2015).
27. Gill, C. D. & O'Brien, T. J. Hubble Space Telescope imaging and ground-based spectroscopy of old nova shells – I. FH Ser, V533 Her, BT Mon, DK Lac, and V476 Cyg. *Mon. Not. R. Astron. Soc.* **314**, 175–182 (2000).
28. Lodders, K., & Amari, S. Presolar grains from meteorites: remnants from the early times of the solar system. *Chem. Erde–Geochem.* **65**, 93–166 (2005).
29. Iliadis, C., Downen, L., José, J., Nittler, L. & Starrfield, S. On presolar stardust grains from CO classical novae. *Astrophys. J.* **855**, 76 (2018).
30. Amari, S. et al. Presolar grains from novae. *Astrophys. J.* **551**, 1065–1072 (2001).

Acknowledgements We thank D. Arnett for insight into stellar evolution modelling and rare-isotope production, including one-dimensional versus three-dimensional simulations. This research was supported by NSF grant AST-1515568 and by NASA under agreement number NNX15AD94G.

Reviewer information *Nature* thanks W. Irvine and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions D.R.S. and L.M.Z. conducted observations of astronomical objects, as well as data reduction and analysis. N.J.W. and T.J.Z. helped in the scientific interpretation of the data with regard to stellar evolution and presolar grains studies, respectively. All authors wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0763-1>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0763-1>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to L.M.Z.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Observations. Observations of all molecular rotational transitions were conducted using the ARO 12-m telescope and SMT, located on Kitt Peak, Arizona and Mt Graham, Arizona, respectively, as well as the IRAM 30-m telescope on Pico Veleta, Spain. The telescopes used for each transition are listed in Extended Data Table 1. Observations made with the ARO 12-m telescope used a dual-polarization receiver that employs ALMA Band 3 sideband-separating mixers. Image rejection was at least 16 dB. The temperature scale (T_A^*) for the 12-m telescope, SMT and 30-m telescope was determined by the chopper-wheel method and is related to the main-beam brightness temperature, T_R , through the equation $T_R = T_A^* / \eta_b$, where η_b is the main-beam efficiency. The 12-m telescope data were obtained using the 1-MHz filterbank and ARO wideband spectrometer (set to a resolution of 0.625 MHz) as backends, both operated in parallel mode to accommodate the two receiver polarizations. The observations conducted with the SMT made use of the 1.3-mm dual-polarization receiver, which contains ALMA Band 6 sideband-separating mixers. Image rejection is at least 15 dB. The 1-MHz filterbank was used as the primary backend. At IRAM, the measurements were made with the Eight Mixer Receiver operating at 2 mm in dual-polarization mode. The FTS 200 (fast Fourier transform spectrometer with a resolution of 200 kHz) and WILMA (wideband line multiple autocorrelator) with a resolution of 2 MHz were employed to acquire the data. All data were fitted with Gaussians at a resolution of 1 MHz; the results of these fits, including the peak antenna temperature, T_A^* , the LSR velocity, V_{LSR} , and the full-width at half-maximum of the line, $\Delta V_{1/2}$, are provided in Extended Data Table 1.

Analysis. The isotope ratios were derived from the molecular column densities of a number of molecules (see Extended Data Table 1 for a full listing of the observed molecular transitions). These ratios are largely consistent with those derived directly from the line intensity ratios, which provide an additional constraint.

The column densities used to determine the given isotope ratios were estimated using the non-LTE radiative transfer code RADEX¹². The program solves the equation of radiative transfer by using the Sobolev escape probability method, thus producing line intensities for comparison with those observed for a given molecule. To do so, RADEX takes an assumed gas kinetic temperature (T_k), H_2 density ($n(H_2)$) and molecular column density (N_{tot}) as input, as well as data files containing energy level, transition and collision rate information for the given molecule; for our purposes, these were obtained from the Leiden Atomic and Molecular Database (LAMDA)³¹. To determine the optimal values for the fitting

parameters (T_k , $n(H_2)$ and N_{tot}), all possible combinations were considered over the ranges 10–60 K for gas kinetic temperature, 1×10^3 – 1×10^7 cm⁻³ for H_2 density and 1×10^{10} – 5×10^{17} for column density. The ‘best fit’ was determined by minimizing the reduced χ^2 . For several molecules, only two rotational transitions were measured; thus, only two parameters could be varied. In these cases the gas kinetic temperature was set to a value accurately determined from other molecules with compatible dipole moments in which multiple transitions were observed. For example, the kinetic temperature derived from five transitions of HC_3N (about 3.7 D) was used for HCN, HNC and their isotopologues (dipole moments of about 3.0–3.1 D).

The measured line intensities (T_A^*) for each molecule (see Extended Data Table 1) were corrected for beam dilution and main-beam efficiency (η_b) for the RADEX analysis. The beam efficiencies for the ARO 12-m telescope, SMT and IRAM 30-m telescope at each transition frequency are given in Extended Data Table 1. The beam-filling factors were determined on the basis of the H_2 map of K4-47 of Akras et al.¹⁰. Extensive mapping of the Helix Nebula in H_2 and HCO^+ showed that the distributions of the two molecules are extremely similar^{32,33}; thus, for K4-47, we assumed that the spread of our observed molecules was comparable to the H_2 distribution imaged by Akras et al.¹⁰. As a check, the beam-filling factor was varied by a factor of two from the nominal value established from the H_2 distribution. The resulting ratios did not vary by more than 8%, except in the cases where the reduced χ^2 values indicated a poor fit. The resulting column densities for each molecule are given in the final column of Extended Data Table 1.

Data availability

All data supporting the findings of this study are available within the paper. Any additional information may be obtained from the corresponding author upon reasonable request.

- Schöier, F. L., van der Tak, F. F. S., van Dishoeck, E. F. & Black, J. H. An atomic and molecular database for analysis of submillimetre line observations. *Astron. Astrophys.* **432**, 369–379 (2005).
- Speck, A. K. et al. Large-scale extended emission around the Helix Nebula: dust, molecules, atoms, and ions. *Astron. J.* **123**, 346–361 (2002).
- Zeigler, N. R., Zack, L. N., Woolf, N. J. & Ziurys, L. M. The Helix Nebula viewed in HCO^+ : large-scale mapping of the $J = 1 \rightarrow 0$ transition. *Astrophys. J.* **778**, 16 (2013).

Extended Data Table 1 | Line parameters for observed molecules

| Molecule | Transition | Frequency (GHz) | Telescope | η_b | $T_A^*{}^a$ (mK) | V_{LSR} (km s ⁻¹) | $\Delta V_{1/2}$ (km s ⁻¹) | N_{tot} (cm ⁻²) |
|-------------------------------|--|-----------------|-----------|----------|------------------|---------------------------------|--|-------------------------------|
| HCN ^b | J=1→0 | 88.6318 | 12-m | 0.67 | 9.0±3.0 | -25.9±6.8 | 39.5±6.8 | 1.5±0.4 × 10 ¹⁴ |
| | J=3→2 | 265.8862 | SMT | 0.75 | 31.0±5.0 | -25.6±2.2 | 39.8±2.2 | |
| H ¹³ CN | J=1→0 | 86.3399 | 12-m | 0.88 | 6.4±0.4 | -23.8±3.5 | 38.0±5.4 | 8.8±0.3 × 10 ¹³ |
| | J=3→2 | 259.0118 | SMT | 0.75 | 26.0±5.5 | -26.5±3.6 | 38.0±4.8 | |
| HC ¹⁵ N | J=1→0 | 86.0550 | 12-m | 0.88 | 1.1±0.3 | -25.3±1.8 | 28.0±7.0 | 1.1±0.5 × 10 ¹³ |
| | J=3→2 | 258.1571 | SMT | 0.75 | 5.5±1.3 | -26.4±3.3 | 30.0±6.0 | |
| HNC ^c | J=1→0 | 90.6636 | 12-m | 0.88 | 2.9±1.0 | -28.4±3.3 | 29.8±3.3 | 3.1±1.2 × 10 ¹³ |
| | J=3→2 | 271.9810 | SMT | 0.75 | 11.5±2.0 | -27.8±2.2 | 36.9±3.3 | |
| HN ¹³ C | J=1→0 | 87.0908 | 12-m | 0.88 | 2.1±0.4 | -23.5±3.5 | 34.0±5.3 | 2.5±0.2 × 10 ¹³ |
| | J=3→2 | 261.2635 | SMT | 0.75 | 4.8±1.5 | -26.0±2.2 | 38.0±3.3 | |
| CN ^d | N,J,F=1, $\frac{3}{2}, \frac{5}{2}$ → 0, $\frac{1}{2}, \frac{3}{2}$ | 113.4909 | 12-m | 0.88 | 3.9±1.5 | -28.2±2.6 | 35.0±2.6 | 2.4±1.0 × 10 ¹⁴ |
| | N,J,F=2, $\frac{3}{2}, \frac{5}{2}$ → 1, $\frac{1}{2}, \frac{3}{2}$ | 226.6595 | SMT | 0.76 | 6.0±2.0 | -27.4±2.6 | 37.0±2.6 | |
| | N,J,F=2, $\frac{3}{2}, \frac{5}{2}$ → 1, $\frac{3}{2}, \frac{5}{2}$ | 226.8742 | SMT | 0.76 | 5.9±2.0 | -27.1±3.9 | 37.0±3.9 | |
| | N,J,F ₁ ,F=2, $\frac{3}{2}, \frac{5}{2}, 2, 3$ → 1, $\frac{1}{2}, 1, 2$ | 217.3032 | SMT | 0.76 | 2.7±0.5 | -27.0±2.8 | 37.0±2.8 | 1.4±0.1 × 10 ¹⁴ |
| ¹³ CN ^d | N,J,F ₁ ,F=2, $\frac{3}{2}, \frac{5}{2}, 3, 4$ → 1, $\frac{3}{2}, 2, 3$ | 217.4652 | SMT | 0.76 | 3.7±0.5 | -27.0±2.8 | 35.0±5.6 | |
| | N,J,F ₁ ,F=2, $\frac{3}{2}, \frac{5}{2}, 2, 3$ → 1, $\frac{3}{2}, 1, 2$ | 217.4286 | SMT | 0.76 | 3.3±0.1 | -27.0±2.8 | 35.0±7.6 | |
| | J=10→9 | 90.9790 | 12-m | 0.88 | 2.0±0.7 | -24.0±3.4 | 35.0±3.3 | 6.4±0.4 × 10 ¹³ |
| | J=15→14 | 136.4644 | IRAM | 0.74 | 25.0±1.5 | -29.0±2.0 | 37.0±4.0 | |
| H ¹³ CCCN | J=17→16 | 154.6573 | IRAM | 0.71 | 27.0±1.5 | -26.7±2.0 | 40.0±4.0 | |
| | J=25→24 | 227.4189 | SMT | 0.76 | 4.3±0.8 | -25.6±3.9 | 37.0±3.9 | |
| | J=26→25 | 236.5128 | SMT | 0.76 | 4.3±1.8 | -28.5±3.3 | 37.0±5.2 | |
| | J=15→14 | 132.2464 | IRAM | 0.75 | 13.0±1.5 | -27.2±2.0 | 35.0±4.0 | 3.1±0.1 × 10 ¹³ |
| HCC ¹³ CN | J=17→16 | 149.8770 | IRAM | 0.72 | 12.5±1.5 | -29.4±2.0 | 37.0±4.0 | |
| | J=15→14 ^e | 135.8986 | IRAM | 0.74 | 14.7±1.5 | -28.0±2.0 | 35.0±4.0 | 3.2±0.2 × 10 ¹³ |
| HC ¹³ CCN | J=17→16 ^e | 154.0161 | IRAM | 0.71 | 17.0±1.5 | -29.0±2.0 | 37.0±4.0 | |
| | J=15→14 ^e | 135.8855 | IRAM | 0.74 | 14.7±1.5 | -28.0±2.0 | 35.0±4.0 | 3.2±0.2 × 10 ¹³ |
| CS ^f | J=17→16 ^e | 154.0012 | IRAM | 0.71 | 17.0±1.5 | -29.0±2.0 | 37.0±4.0 | |
| | J=2→1 | 97.9810 | 12-m | 0.64 | 1.0±0.5 | -27.3±3.5 | 19.8±6.1 | 1.8±1.0 × 10 ¹³ |
| | J=3→2 | 146.9690 | 12-m | 0.50 | 2.0±1.0 | -23.0±2.6 | 19.1±2.7 | |
| ¹³ CS | J=5→4 | 244.9356 | SMT | 0.76 | 2.0±0.7 | -28.0±2.4 | 22.3±1.2 | |
| | J=3→2 | 138.7393 | IRAM | 0.74 | 2.9±1.5 | -25.1±2.0 | 23.0±4.0 | 3.9±1.6 × 10 ¹² |
| CO ^f | J=1→0 | 115.2712 | 12-m | 0.63 | 37.0±7.0 | -25.8±2.1 | 19.6±3.0 | 2.1±0.1 × 10 ¹⁷ |
| | J=2→1 | 230.5380 | SMT | 0.76 | 179.0±4.0 | -26.2±1.0 | 19.6±2.1 | |
| | J=3→2 | 345.7960 | SMT | 0.66 | 140.0±12.0 | -26.5±1.0 | 23.0±2.1 | |
| | J=6→5 | 691.4731 | SMT | 0.60 | 41.0±17.0 | -23.4±1.0 | 22.6±2.4 | |
| C ¹⁷ O | J=1→0 | 112.3593 | 12-m | 0.88 | 4.2±1.3 | -25.0±1.4 | 21.6±5.4 | 9.8±4.7 × 10 ¹⁵ |
| | J=2→1 | 224.7144 | SMT | 0.76 | 15.5±6.3 | -25.0±2.2 | 21.0±3.3 | |

^aMeasured with a resolution of 1 MHz.^bFrom Schmidt & Ziurys¹³.^cFrom Schmidt & Ziurys¹⁴.^dStrongest hyperfine component only.^eBlended lines.^fFrom Edwards et al.¹¹.

Impact of pear-shaped fission fragments on mass-asymmetric fission in actinides

Guillaume Scamps^{1*} & Cédric Simenel^{2,3}

Nuclear fission of heavy (actinide) nuclei results predominantly in asymmetric mass splits¹. Without quantum shell effects, which can give extra binding energy to their mass-asymmetric shapes, these nuclei would fission symmetrically. The strongest shell effects appear in spherical nuclei, such as the spherical ‘doubly magic’ (that is, both its atomic and neutron numbers are ‘magic’ numbers) nucleus ¹³²Sn, which contains 50 protons and 82 neutrons. However, a systematic study of fission² has shown that heavy fission fragments have atomic numbers distributed around $Z = 52$ to $Z = 56$, indicating that the strong shell effects in ¹³²Sn are not the only factor affecting actinide fission. Reconciling the strong spherical shell effects at $Z = 50$ with the different Z values of fission fragments observed in nature has been a longstanding puzzle³. Here we show that the final mass asymmetry of the fragments is also determined by the extra stability provided by octupole (pear-shaped) deformations, which have been recently confirmed experimentally around ¹⁴⁴Ba ($Z = 56$)^{4,5}, one of very few nuclei with shell-stabilized octupole deformation⁶. Using a quantum many-body model of superfluid fission dynamics⁷, we find that heavy fission fragments are produced predominantly with 52 to 56 protons, which is associated with substantial octupole deformation acquired on the way to fission. These octupole shapes, which favour asymmetric fission, are induced by deformed shells at $Z = 52$ and $Z = 56$. By contrast, spherical magic nuclei are very resistant to octupole deformation, which hinders their production as fission fragments. These findings may explain surprising observations of asymmetric fission in nuclei lighter than lead⁸.

Atomic nuclei are usually found at a minimum of energy, the ground state, which may be deformed because of quantum correlations. Elongation beyond the ground state costs potential energy, until a maximum is reached at the fission barrier. Increasing the elongation beyond the fission barrier decreases the potential energy, and the system follows a ‘fission valley’ in the potential energy surface until it breaks into two fragments (scission). In the absence of quantum shell effects, all heavy nuclei preferentially fission into two fragments of similar mass (mass-symmetric fission). However, quantum shell effects in the fissioning nucleus can result in several valleys on the way to scission. These may be mass-symmetric or mass-asymmetric.

Although progress has been made recently in describing fission-fragment mass distributions with stochastic approaches^{9,10}, theoretical description of the first stage of fission, from the ground-state deformation to the fission barrier, remains a challenge¹¹. However, the study of the dynamics along the fission valleys is now possible with the time-dependent energy-density functional approach^{12,13} including nuclear superfluidity^{7,14,15}. This approach is quantum-based and fully microscopic in that it follows the evolution of all single-particle wavefunctions (196 for neutrons and 126 for protons in the present calculations), which are fully or partly occupied, in time.

Figure 1 shows the dynamical evolution of isodensity surfaces of ²⁴⁰Pu, starting from a configuration in the asymmetric-fission valley. The final state corresponds to a quantum superposition of different repartitions of the number of nucleons between the fragments, with $\langle Z \rangle \approx 53.8$

protons and $\langle N \rangle \approx 85.2$ neutrons in average in the heavy (left) fragment. Similar calculations were performed for the actinides ²³⁰Th, ^{234,236}U, ²⁴⁶Cm, ²⁵⁰Cf and ²⁵⁸Fm (Extended Data Table 1 and Extended Data Fig. 1). For each system, a range of initial configurations in the asymmetric fission valley has been considered to investigate the influence of the initial elongation (and consequently the initial potential energy) on the final properties of the fragments. In the example of Fig. 1, scission occurs after about 20 zs (1 zs = 10^{-21} s), but can reach up to 90 zs depending on the initial configuration. Despite large fluctuations in the time to scission, almost all heavy fragments are formed with $Z = 52$ – 56 protons (see Fig. 2), in excellent agreement with experimental data. The fact that the number of protons shows little (if any) dependence on the initial elongation is due to the slow viscous motion of fissioning nuclei, which essentially follow the bottom of their fission valley¹⁶, except near scission, where the evolution is expected to be faster¹².

We also observe in Fig. 1 and Extended Data Fig. 1 that the fragments are formed with a strong deformation at scission. This deformation results from the competition between the long-range Coulomb interaction, which repels the fragments, the short range nuclear attraction in the neck between the fragments, and the deformation energy of the fragments. The latter quantifies the energy cost to deform a fragment, which can be particularly large for spherical doubly magic nuclei, such as ¹³²Sn, or small for non-magic nuclei, which are often deformed in their ground state.

The strong attractive nuclear force between the fragments is responsible for the neck (see Fig. 1b, c), inducing quadrupole (cigar-shaped) and octupole (pear-shaped) deformations of the fragments. Although quadrupole deformation is often taken into account in modelling fission¹⁷, octupole deformation is also important for describing scission configurations properly¹⁸. The neutron localization function¹⁹ C_n (see Methods), shown as projections in Fig. 1, also exhibits strong octupole shapes. The localization function is often used to characterize shell structures in quantum many-body systems such as nuclei^{20,21} and atoms²¹. For instance, we see in Fig. 1b that the signature shell structures of particular fragments, as well as their deformation, are already present about 4 zs before scission. Examples of identification of the pre-fragment as octupole-deformed ¹⁴⁴Ba ($Z = 56$) are given in Extended Data Figs. 3, 4.

Our calculations show that, indeed, Sn fragments produced in symmetric fission of ²⁵⁸Fm have octupole moments 2 to 3 times smaller at scission than those of fragments with $Z \approx 55$. Fragments with $Z \approx 52$ also exhibit substantial octupole deformation, although not as strong as that of fragments with $Z \approx 55$ (see Extended Data Fig. 2). The case of the spherical, doubly magic ¹³²Sn nucleus is particularly interesting. On the one hand, its production as a fission fragment is usually expected to be favoured because of its extra binding energy, which originates from spherical shell effects. On the other hand, its deformation (which is inevitable in fission) costs more energy, thus hindering its production as a fission fragment.

A similar interplay between intrinsic deformation of the fragments and their relative motion is well known in heavy-ion fusion²². The

¹Center for Computational Sciences, University of Tsukuba, Tsukuba, Japan. ²Department of Nuclear Physics, Research School of Physics and Engineering, Australian National University, Canberra, Australian Capital Territory, Australia. ³Department of Theoretical Physics, Research School of Physics and Engineering, Australian National University, Canberra, Australian Capital Territory, Australia. *e-mail: scamps@nucl.ph.tsukuba.ac.jp

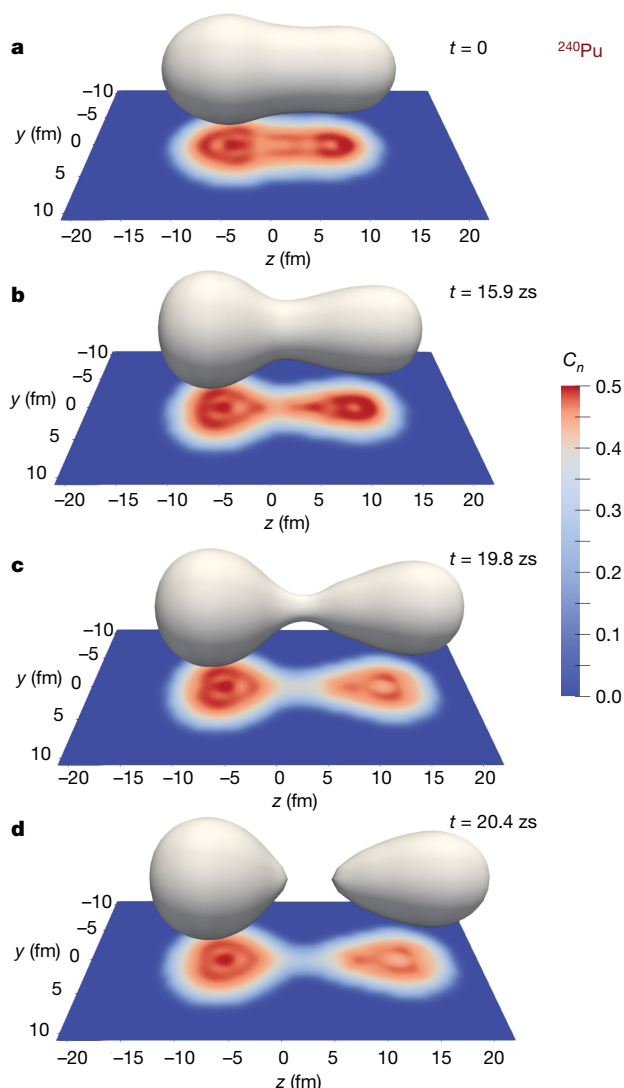


Fig. 1 | Microscopic calculations of asymmetric fission of ^{240}Pu . **a–d**, Isodensity surfaces at 0.08 fm^{-3} (half the saturation density), computed from the full microscopic evolution, are shown at different times (see Supplementary Information video). The localization function C_n of the neutrons (see Methods) is shown in the projections. Here, scission occurs at $t \approx 20 \text{ zs}$, that is, between **c** and **d**. The quadrupole and octupole deformation parameters (see Methods) at scission are $\beta_2 \approx 0.16$ and $\beta_3 \approx 0.22$ for the heavy fragment (left) and $\beta_2 \approx 0.64$ and $\beta_3 \approx 0.4$ for the light fragment (right), respectively.

capture mechanism can naively be seen as the ‘reverse’ process to scission, which occurs at the late stage of nuclear fission. Therefore, it is not surprising that similar couplings to octupole shapes impact fission dynamics, favouring the formation of fission fragments that exhibit octupole correlations. Of course, a similar role is played by quadrupole couplings (see, for instance, the strong quadrupole deformation of the light fragment in Fig. 1). However, as the majority of nuclei exhibit quadrupole deformation in their ground state, this cannot be the only reason for the specific number of protons ($Z \approx 52\text{--}56$) of the heavy fragment in asymmetric fission of actinides.

By contrast, fewer nuclei are expected to exhibit octupole deformation in their ground state^{4,6,23–25}. Recent experiments have confirmed non-ambiguously that this is the case for ^{144}Ba ($Z = 56$)⁴, a possible heavy fragment in asymmetric fission of actinides. Nuclei close to ^{144}Ba in the nuclear chart should also exhibit particularly strong octupole correlations, thereby providing a possible explanation to the favoured production of these nuclei in fission. The microscopic approach used here automatically incorporates the possibility for the nuclei to acquire

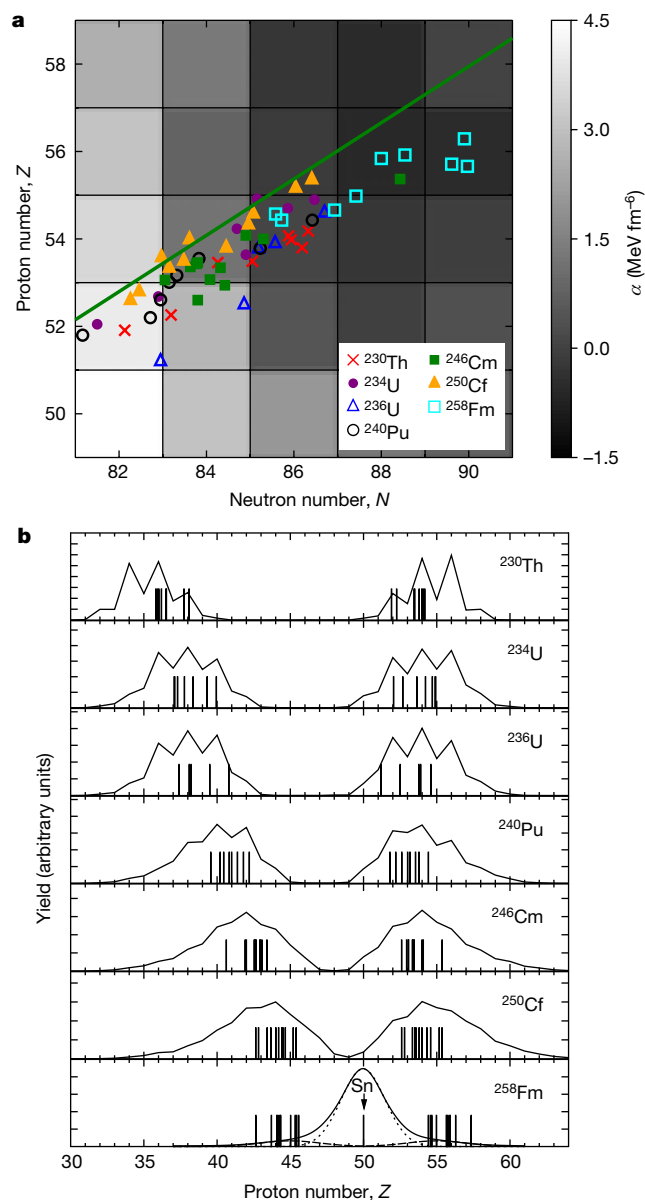


Fig. 2 | Proton and neutron number distributions in fission fragments.

a, Expectation values of the number of protons (Z) and neutrons (N) in the heavy fragments for various asymmetric fissions. The solid line shows the expected positions of fragments with the N/Z values of ^{240}Pu . The background grey scale quantifies the resistance to octupole (pear-shaped) deformations in the nuclei, as predicted by constrained Hartree–Fock calculations with dynamical Bardeen–Cooper–Schrieffer pairing correlations. It is obtained from the curvature of the octupole deformation energy $\alpha = \lim_{Q_{30} \rightarrow 0} (E/Q_{30}^2)$ of the nuclei, where Q_{30} is the octupole moment (see Methods), near their energy minimum at $Q_{30} = 0$ (which corresponds to the curvature at the origin in Fig. 3b). Negative values indicate nuclei likely to exhibit octupole deformations in their ground state. **b**, Our microscopic predictions of the expectation values $\langle Z \rangle$ of the number of protons in the fission fragments (vertical lines) are compared with fragment proton number distributions (solid lines) extracted from experimental results of thermal-neutron-induced fission (top six panels)³⁰ and the spontaneous fission of ^{258}Fm (bottom panel)²⁶. Dashed and dotted lines in the bottom panel are Gaussian fits of the asymmetric and symmetric components, respectively.

octupole deformations induced by their underlying quantum shell structure. This is illustrated in Fig. 3a, which shows the potential energy surface of ^{144}Ba and predicts the ground state to be quadrupole- and octupole-deformed. According to the deformation energies plotted in Fig. 3b, these octupole correlations are present in nuclei with proton

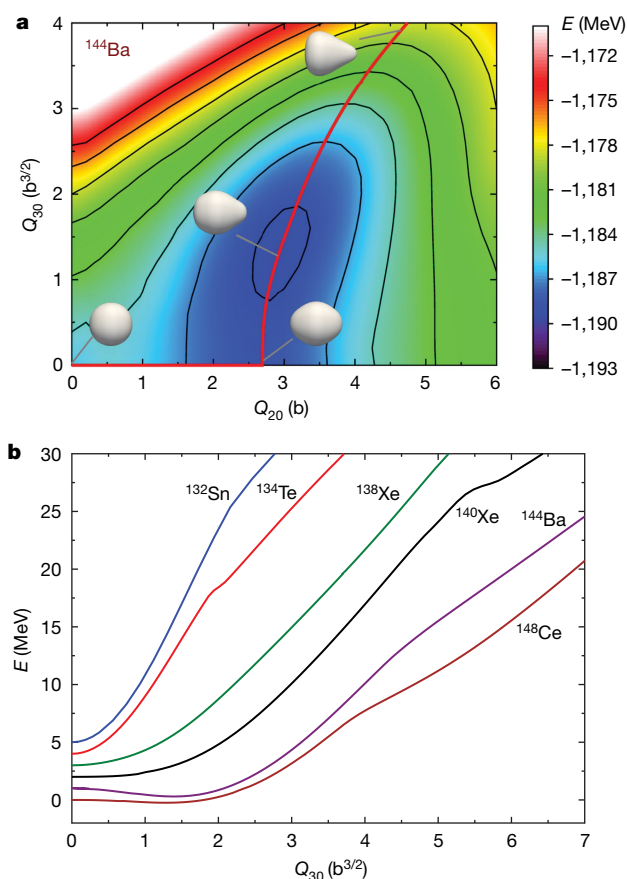


Fig. 3 | Deformation energy in fission fragments. a, Potential energy surface of ^{144}Ba . The binding energy, obtained using the Skyrme SLy4d functional³¹, is shown as a function of the quadrupole (Q_{20}) and octupole (Q_{30}) moments (see Methods). The binding energy increases from blue to red, with the isoenergy contour lines separated by an energy of 2 MeV. The red thick line represents the minimum energy at a given quadrupole deformation. **b**, Microscopic calculations of octupole deformation energy (defined as the binding energy of an octupole deformed nucleus minus the binding energy of the same nucleus without octupole deformation) for several isotopes produced in the fission of actinides. All other multipole moments, including the quadrupole moment, are not constrained. Thus, the curves show the minimum energy for a given octupole moment. The reference energy is shifted up by 1-MeV steps for each curve for clarity. Whereas ^{144}Ba is predicted to have an octupole minimum, these octupole correlations disappear for the magic nucleus Sn ($Z = 50$).

and neutron numbers close to those of ^{144}Ba and ^{140}Xe . However, they disappear in nuclei close to ^{132}Sn which, as expected, are found to be resistant against octupole deformations (see also Extended Data Fig. 5).

The origin of these octupole correlations can be understood from the shell structure of the nuclei. Like in atoms, the properties of the spectrum of single-particle energy levels are often found to be responsible for quantum shell effects in atomic nuclei. For instance, spherical nuclei with a magic number of protons and neutrons (the analogue of noble gases) have all their proton and neutron levels filled below large energy gaps of a few megaelectronvolts, above which the levels are empty. As a result, the nucleus is difficult to excite and acquires an extra binding energy from these quantum shell effects. However, the single-particle energy spectrum changes with the deformation of the nucleus in such a way that spherical energy gaps disappear, while other gaps may appear in deformed nuclei, stabilizing their shape. The calculated neutron and proton single-particle energies in ^{144}Ba are shown in Fig. 4 as a function of quadrupole (Q_{20}) and octupole (Q_{30}) deformation, following the path of minimum energy (solid red line in Fig. 3a). The large energy gaps at the spherical point ($Q_{20} = Q_{30} = 0$) correspond to the magic numbers $Z = 50$ and $N = 82$, which are responsible for the shell effects

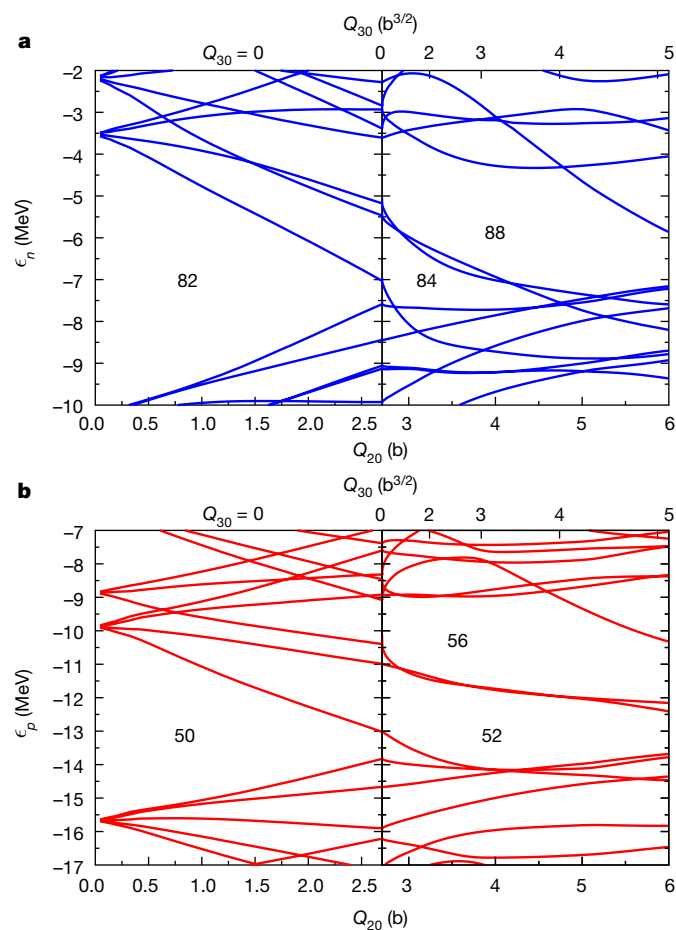


Fig. 4 | Evolution of single-particle energies with deformation.

a, b, Neutron (ϵ_n ; **a**) and proton (ϵ_p ; **b**) single-particle energies as a function of the quadrupole (Q_{20} ; lower scale) and octupole (Q_{30} ; upper scale) moments in ^{144}Ba following the path of the red solid line in Fig. 3a. The numbers in the energy gaps correspond to the number of particles that can occupy the single-particle states below the energy gap. 50 and 82 are magic numbers, associated with spherical shells, whereas 52, 56, 84 and 88 are associated with deformed energy gaps. In particular, the opening of the $Z = 56$ and $N = 88$ energy gaps with octupole deformation induces collective octupole correlations in the ground state of nuclei around ^{144}Ba . Evidence for such correlations³² includes additional binding, revealed by measured masses in this region, as well as shallow minima in theoretical potential energy surfaces for octupole deformed shapes, as shown in Fig. 3a. However, direct experimental evidence for octupole deformation in this region was only recently found^{4,5}.

in ^{132}Sn . However, we observe a closure of these spherical energy gaps and the opening of deformed shell gaps at $Z = 52, 56$ and $N = 84, 88$, which survive for a large range of octupole deformations. The matching between the positions of the deformed proton shell gaps $Z = 52, 56$ in Fig. 4b and the proton numbers of the heavy fragments (see Fig. 2b) is striking. The fact that the final mass asymmetry can be explained by octupole-deformed shells without invoking the $Z = 50$ spherical gap is surprising. We note that the observed importance of octupole correlations at scission does not exclude a possible contribution of magic shells at the early stage of fission, for example, via the creation of valleys in the potential energy surface. Several competing effects are then expected to be at play in the formation of these valleys. These include the spherical and octupole shell gaps of the heavy fragment, in addition to other possible deformed shell gaps in the light fragment. In fact, amongst all actinides studied experimentally, only ^{258}Fm is clearly dominated by $Z = 50$ spherical shell effects²⁶, producing two symmetric Sn fragments responsible for the narrow peak in the fission fragment charge distribution of ^{258}Fm (see Fig. 2b), although it also exhibits a weaker

asymmetric mode with a heavy fragment around $Z \approx 55$. However, asymmetric fission dominates in ^{256}Fm and lighter fermium isotopes²⁷, confirming the weak influence of $Z = 50$ in the formation of fission fragments.

The total kinetic energy (TKE) of the final fragments is another important observable that can be used to characterize fission properties. It is related to the elongation of the system at scission and thus to the deformation of the fragments. For instance, a large elongation at scission reduces the Coulomb repulsion between the fragments, therefore leading to a smaller TKE when the fragments are far from each other and the Coulomb energy has been converted into kinetic energy. By contrast, a compact system with near spherical fragments at scission is usually associated to larger TKE. The TKE computed from our microscopic simulations, which follow the method introduced in ref.¹², are in very good agreement with experimental data^{28,29} (see Extended Data Table 1 and Extended Data Fig. 6). In particular, most of our calculated TKEs are found to be smaller for fission leading to $Z \approx 55$ fragments than that producing $Z \approx 52$ fragments, indicating that in the latter case the fissioning nucleus is more compact at scission. The smaller TKE for $Z \approx 55$ is interpreted as an effect of the larger deformation in the heavy fragments.

The fact that both the mass-asymmetric fission of actinides and the TKE measured experimentally can be explained by our time-dependent microscopic calculations gives us confidence in the fission dynamics predicted by these calculations and, in particular, in the major role of the octupole deformation of the heavy fragments. The octupole deformation in $Z_{\text{light}} \approx 34$ and $N_{\text{heavy}} \approx 56$ nuclei²⁵ could also explain mass-asymmetric fission found experimentally⁸ in lighter systems. Other properties of fission will also be investigated in the future, such as the excitation energy of the fragments and the number of neutrons emitted during the fission process.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0780-0>.

Received: 27 April 2018; Accepted: 31 October 2018;

Published online 19 December 2018.

- Andreyev, A. N., Nishio, K. & Schmidt, K.-H. Nuclear fission: a review of experimental advances and phenomenology. *Rep. Prog. Phys.* **81**, 016301 (2018).
- Schmidt, K.-H. et al. Relativistic radioactive beams: a new access to nuclear-fission studies. *Nucl. Phys. A* **665**, 221–267 (2000).
- Schmidt, K.-H. & Jurado, B. Review on the progress in nuclear fission – experimental methods and theoretical descriptions. *Rep. Prog. Phys.* **81**, 106301 (2018).
- Bucher, B. et al. Direct evidence of octupole deformation in neutron-rich ^{144}Ba . *Phys. Rev. Lett.* **116**, 112503 (2016).
- Bucher, B. et al. Direct evidence for octupole deformation in ^{146}Ba and the origin of large $E1$ moment variations in reflection-asymmetric nuclei. *Phys. Rev. Lett.* **118**, 152504 (2017).
- Gaffney, L. P. et al. Studies of pear-shaped nuclei using accelerated radioactive beams. *Nature* **497**, 199–204 (2013).
- Scamps, G., Simenel, C. & Lacroix, D. Superfluid dynamics of ^{258}Fm fission. *Phys. Rev. C* **92**, 011602 (2015).
- Andreyev, A. N. et al. New type of asymmetric fission in proton-rich nuclei. *Phys. Rev. Lett.* **105**, 252502 (2010).
- Möller, P. & Randrup, J. Calculated fission-fragment yield systematics in the region $74 \leq Z \leq 94$ and $90 \leq N \leq 150$. *Phys. Rev. C* **91**, 044316 (2015).
- Sadhukhan, J., Nazarewicz, W. & Schunck, N. Microscopic modeling of mass and charge distributions in the spontaneous fission of ^{240}Pu . *Phys. Rev. C* **93**, 011304 (2016).
- Schunck, N. & Robledo, L. M. Microscopic theory of nuclear fission: a review. *Rep. Prog. Phys.* **79**, 116301 (2016).
- Simenel, C. & Umar, A. S. Formation and dynamics of fission fragments. *Phys. Rev. C* **89**, 031601 (2014).
- Goddard, P. M., Stevenson, P. D. & Rios, A. Fission dynamics within time-dependent Hartree–Fock: deformation-induced fission. *Phys. Rev. C* **92**, 054610 (2015).
- Bulgac, A., Magierski, P., Roche, K. J. & Stetcu, I. Induced fission of ^{240}Pu within a real-time microscopic framework. *Phys. Rev. Lett.* **116**, 122504 (2016).
- Tanimura, Y., Lacroix, D. & Ayik, S. Microscopic phase-space exploration modeling of ^{258}Fm spontaneous fission. *Phys. Rev. Lett.* **118**, 152501 (2017).
- Bulgac, A., Jin, S., Roche, K., Schunck, N. & Stetcu, I. Fission dynamics. Preprint at <https://arxiv.org/abs/1806.00694> (2018).
- Möller, P., Madland, D. G., Sierk, A. J. & Iwamoto, A. Nuclear fission modes and fragment mass asymmetries in a five-dimensional deformation space. *Nature* **409**, 785–790 (2001).
- Carjan, N., Ivanyuk, F. A. & Oganessian, Y. T. Pre-scission model predictions of fission fragment mass distributions for super-heavy elements. *Nucl. Phys. A* **968**, 453–464 (2017).
- Becke, A. D. & Edgecombe, K. E. A simple measure of electron localization in atomic and molecular systems. *J. Chem. Phys.* **92**, 5397–5403 (1990).
- Reinhard, P.-G., Maruhn, J. A., Umar, A. S. & Oberacker, V. E. Localization in light nuclei. *Phys. Rev. C* **83**, 034312 (2011).
- Jerabek, P., Schuettrumpf, B., Schwerdtfeger, P. & Nazarewicz, W. Electron and nucleon localization functions of oganesson: approaching the Thomas–Fermi limit. *Phys. Rev. Lett.* **120**, 053001 (2018).
- Dasgupta, M., Hinde, D. J., Rowley, N. & Stefanini, A. M. Measuring barriers to fusion. *Annu. Rev. Nucl. Part. Sci.* **48**, 401–461 (1998).
- Butler, P. & Nazarewicz, W. Intrinsic reflection asymmetry in atomic nuclei. *Rev. Mod. Phys.* **68**, 349–421 (1996).
- Robledo, L. M. & Bertsch, G. F. Global systematics of octupole excitations in even-even nuclei. *Phys. Rev. C* **84**, 054302 (2011).
- Butler, P. A. Octupole collectivity in nuclei. *J. Phys. G* **43**, 073002 (2016).
- Hulet, E. K. et al. Bimodal symmetric fission observed in the heaviest elements. *Phys. Rev. Lett.* **56**, 313–316 (1986).
- Unik, J. P., Glendenin, L. E., Flynn, K. F., Gorski, A. & Sjöblom, R. K. Fragment mass and kinetic energy distributions for fissioning systems ranging from mass 230 to 256. In *Proc. of the Third International Atomic Energy Agency Symposium on the Physics and Chemistry of Fission*, Vol. II 19–45 (IAEA, 1974).
- Böckstiegel, C. et al. Nuclear-fission studies with relativistic secondary beams: Analysis of fission channels. *Nucl. Phys. A* **802**, 12–25 (2008).
- Caamaño, M. et al. Characterization of the scission point from fission-fragment velocities. *Phys. Rev. C* **92**, 034606 (2015).
- Brown, D. A. et al. ENDF/B-VIII.0: the 8th major release of the nuclear reaction data library with cello-project cross sections, new standards and thermal scattering data. *Nucl. Data Sheets* **148**, 1–142 (2018).
- Kim, K.-H., Otsuka, T. & Bonche, P. Three-dimensional TDHF calculations for reactions of unstable nuclei. *J. Phys. G Nucl. Phys.* **23**, 1267–1273 (1997).
- Leander, G. A., Nazarewicz, W., Olanders, P., Ragnarsson, I. & Dudek, J. A new region of intrinsic reflection asymmetry in nuclei around ^{145}Ba ? *Phys. Lett. B* **152**, 284–290 (1985).

Acknowledgements We thank B. Jurado, A. Chatillon and F. Farget for useful discussions at the early stage of this work. We are grateful to D. J. Hinde for continuous support to this project. We thank M. Caamaño for providing references to experimental data. B. Jurado and D. J. Hinde are also thanked for their careful reading of the manuscript. This work has been supported by the Australian Research Council under grant number DP160101254. The calculations were performed in part at the NCI National Facility in Canberra, Australia, which is supported by the Australian Commonwealth Government, in part using the COMA system at the CCS in the University of Tsukuba, which is supported by the HPCI Systems Research Projects (project hp180041), and using the Oakforest-PACS at the JCAHPC in Tokyo, which is supported in part by the Multidisciplinary Cooperative Research Program in CCS, University of Tsukuba.

Author contributions G.S. and C.S. conceived the project. G.S. performed the numerical simulations. G.S. and C.S. discussed the results. C.S. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0780-0>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0780-0>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to G.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Microscopic calculations with pairing correlations. Constrained and time-dependent Hartree-Fock calculations with dynamical Bardeen-Cooper-Schrieffer pairing correlations (CHF+BCS and TDBCS, respectively) were done in a three-dimensional Cartesian geometry with one plane ($y = 0$) of symmetry using the code of ref. ⁷ and the Skyrme SLy4d energy density functional³¹ with a surface pairing interaction of strength $V_0^{nn} = 1,256$ MeV fm³ and $V_0^{pp} = 1,462$ MeV fm³ in the neutron- and proton-pairing channels, respectively³³. The effect of the choice of pairing interaction and functional on the octupole deformation is shown in Extended Data Fig. 7. The cut-off function and parameters are the same as in ref. ³³. The three-dimensional Poisson equation for the Coulomb potential is solved and the Slater approximation is used for the Coulomb exchange term. The TDBCS calculations are first performed in a spatial grid of dimensions $L_x \times 2L_y \times L_z = 19.2 \times 19.2 \times 40$ fm³ until the fragments reach a relative distance of 17 fm between their centres of mass. Then, the system is described in a larger box with $L_z = 56$ fm. The mesh spacing is 0.8 fm in all directions and the time step between two time iterations is 1.5×10^{-24} s.

Following the technique of ref. ⁷, the initial configurations for the time-dependent calculations are generated by CHF+BCS calculations with a combination of quadrupole and octupole constraints used to control the elongation and the mass asymmetry, respectively. Fission occurs along the z axis. The other multipole moments are not constrained. Once in the asymmetric fission valley, the octupole constraint is released to allow the system to find a local minimum of energy for a given quadrupole deformation⁷. A range of 5 to 11 initial quadrupole moments, between $Q_{20} \approx 34$ b and 72 b, is considered for each system with steps of 1–10 b ($1 \text{ b} = 10^{-28} \text{ m}^2$; see Extended Data Table 1).

The CHF+BCS calculations for Fig. 2a (background), 3, 4 and Extended Data Fig. 5 are done with a modified version of the ev8 code³⁴, where only one plane of symmetry ($y = 0$) is used. The spatial grid used for those calculations has dimensions $L_x \times 2L_y \times L_z = 22.4 \times 22.4 \times 25.6$ fm³.

Pairing correlations are treated dynamically, that is, without the frozen occupation approximation. Although, the BCS approximation violates the continuity equation³⁵, the spurious transfer of particles after scission turns out to be very small in the case of fission¹⁵. Compared to the more general Bogoliubov treatment of pairing correlations, the BCS approximation has the advantage of reducing computational needs substantially, whereas the resulting fission dynamics (saddle-to-scission times) and properties of the fission fragments (mass, charge and TKE) are very similar to both Bogoliubov^{14,16} and BCS dynamical pairing^{7,15} results (see also ref. ³⁶ for a review).

Multipole moments and deformation parameters. The quadrupole moment is expressed as $Q_{20} = \sqrt{\frac{5}{16\pi}} \int \rho(\mathbf{r}) (2z^2 - x^2 - y^2) d^3r$ and the octupole moment as $Q_{30} = \sqrt{\frac{7}{16\pi}} \int \rho(\mathbf{r}) [2z^3 - 3z(x^2 + y^2)] d^3r$, where $\rho(\mathbf{r})$ is the density of nucleons. The deformation parameters β_2 and β_3 are obtained from the quadrupole and octupole moments following

$$\beta_\lambda = \frac{4\pi}{3A(r_0 A^{1/3})^\lambda} Q_{\lambda 0} \quad (1)$$

with $r_0 = 1.2$ fm.

Fermion localization function. The localization function is computed as¹⁹

$$C_{q\sigma}(\mathbf{r}) = \left[1 + \left(\frac{\tau_{q\sigma} \rho_{q\sigma} - \frac{1}{4} |\nabla \rho_{q\sigma}|^2 - j_{q\sigma}^2}{\rho_{q\sigma} \tau_{q\sigma}^{\text{TF}}} \right)^2 \right]^{-1} \quad (2)$$

with the nucleon ($\rho_{q\sigma}$), kinetic ($\tau_{q\sigma}$) and current ($j_{q\sigma}$) densities defined as

$$\rho_{q\sigma}(\mathbf{r}) = \sum_{a \in q} n_a \varphi_a^*(\mathbf{r}\sigma) \varphi_a(\mathbf{r}\sigma) \quad (3)$$

$$\tau_{q\sigma}(\mathbf{r}) = \sum_{a \in q} n_a |\nabla \varphi_a(\mathbf{r}\sigma)|^2 \quad (4)$$

$$j_{q\sigma}(\mathbf{r}) = \sum_{a \in q} n_a \text{Im}[\varphi_a^*(\mathbf{r}\sigma) \nabla \varphi_a(\mathbf{r}\sigma)] \quad (5)$$

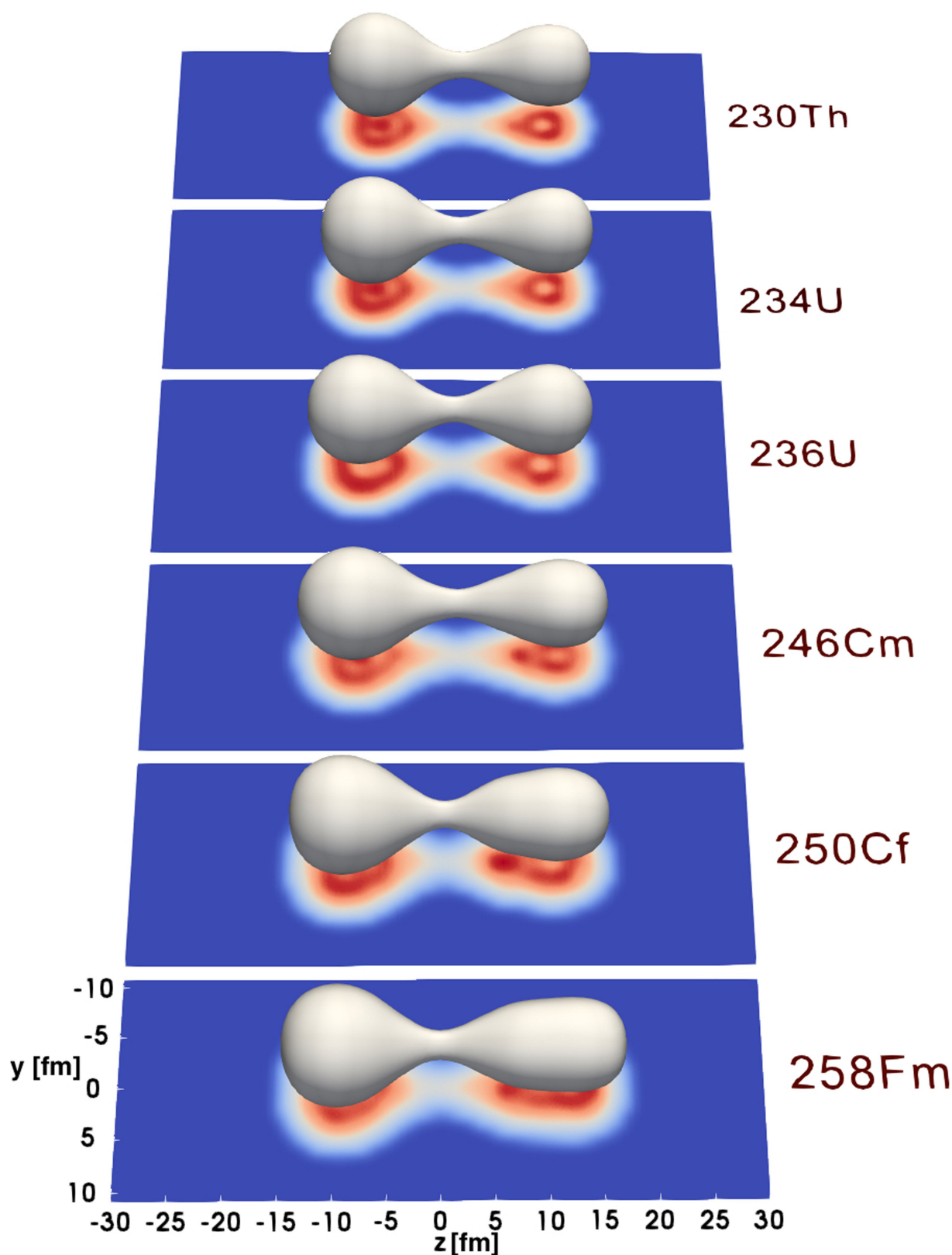
where q stands for neutron (n) or proton (p), σ is the spin, φ are the single-particle wave functions and φ^* are their complex conjugates. τ^{TF} is the Thomas-Fermi approximation of the kinetic density. To study the inner core of the nuclei, we suppress the localization function on the surface of the fragments by applying the transformation³⁷

$$C_{q\sigma}(\mathbf{r}) \rightarrow C_{q\sigma}(\mathbf{r}) \frac{\rho_{q\sigma}(\mathbf{r})}{\max[\rho_{q\sigma}(\mathbf{r})]} \quad (6)$$

The neutron ($q = n$) localization function is obtained by averaging over the spin σ .

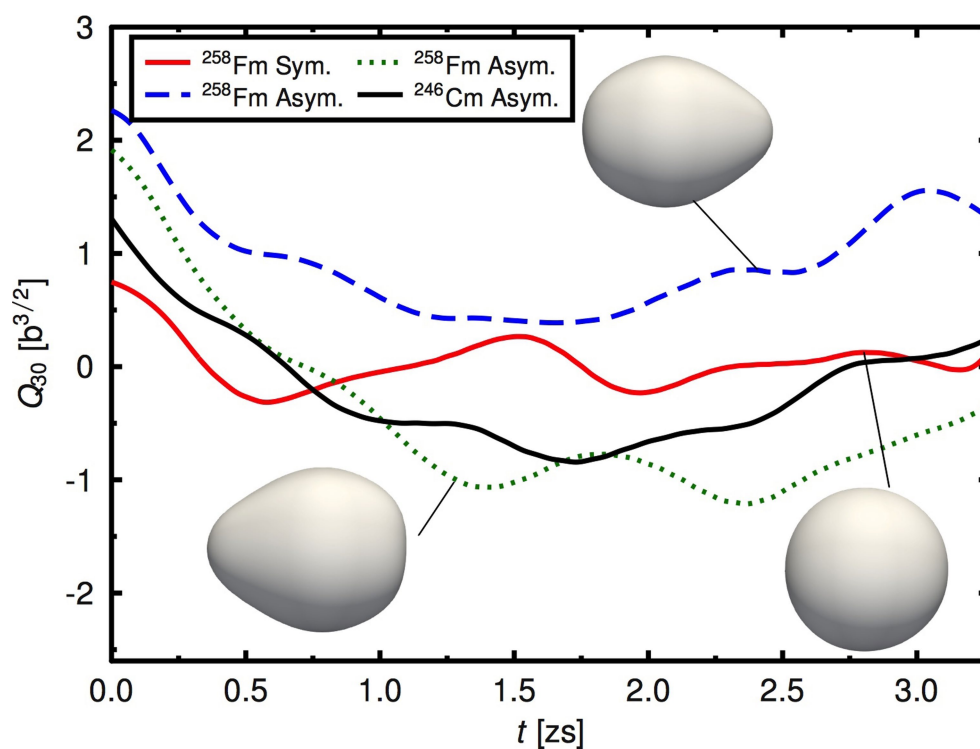
Data and code availability. The datasets generated during the current study and the associated codes are available from the corresponding author on reasonable request.

33. Scamps, G. & Lacroix, D. Effect of pairing on one- and two-nucleon transfer below the Coulomb barrier: A time-dependent microscopic description. *Phys. Rev. C* **87**, 014605 (2013).
34. Bonche, P., Flocard, H. & Heenen, P. H. Solution of the Skyrme HF+BCS equation on a 3D mesh. *Comput. Phys. Commun.* **171**, 49–62 (2005).
35. Scamps, G., Lacroix, D., Bertsch, G. F. & Washiyama, K. Pairing dynamics in particle transport. *Phys. Rev. C* **85**, 034328 (2012).
36. Simenel, C. & Umar, A. S. Heavy-ion collisions and fission dynamics with the time-dependent Hartree-Fock theory and its extensions. *Prog. Part. Nucl. Phys.* **103**, 19–66 (2018).
37. Zhang, C. L., Schuettrumpf, B. & Nazarewicz, W. Nucleon localization and fragment formation in nuclear fission. *Phys. Rev. C* **94**, 064323 (2016).
38. Sadhukhan, J., Zhang, C., Nazarewicz, W. & Schunck, N. Formation and distribution of fragments in the spontaneous fission of ²⁴⁰Pu. *Phys. Rev. C* **96**, 061301 (2017).
39. Warda, M., Staszczak, A. & Nazarewicz, W. Fission modes of mercury isotopes. *Phys. Rev. C* **86**, 024601 (2012).
40. Wilkins, B. D., Steinberg, E. P. & Chasman, R. R. Scission-point model of nuclear fission based on deformed-shell effects. *Phys. Rev. C* **14**, 1832 (1976).
41. Böckstiegel, C. *Bestimmung der Totalen Kinetischen Energien in der Niederenergiespaltung Neutronenarmer Radioaktiver Isotope*. PhD thesis, TU Darmstadt (1998).



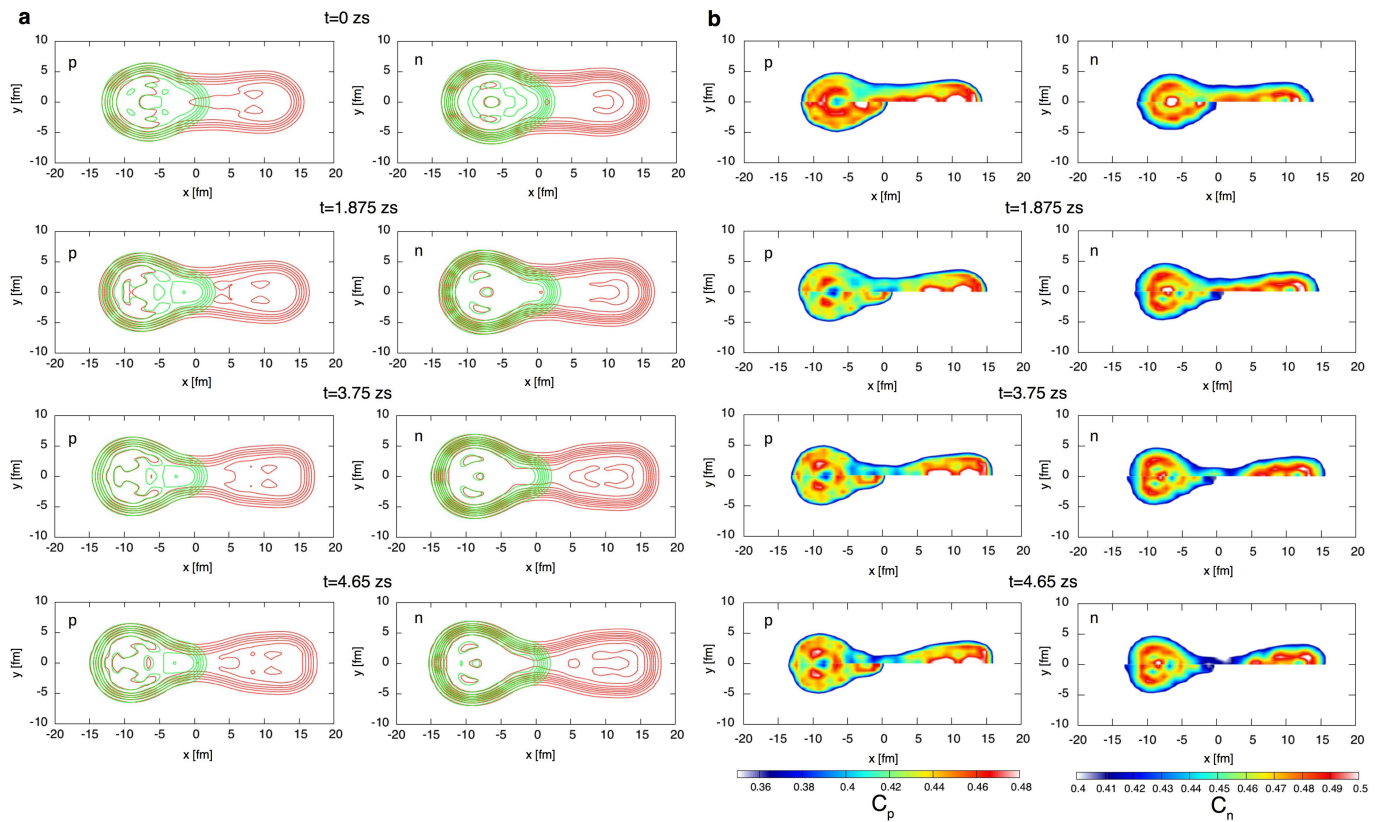
Extended Data Fig. 1 | Scission configurations. Isodensity surface and neutron localization just before scission (about 0.1 zs before the neck breaks), used for calculations with different actinides in their asymmetric fission valleys. At scission, all the heavy fragments (left) have octupole deformation parameters (see Methods) $\beta_3 \approx 0.23$ – 0.27 and $\beta_2 \approx 0.15$ – 0.27 . These fragments are much more deformed than those

produced by symmetric fission of ^{258}Fm (see Extended Data Fig. 2), where symmetric Sn fragments are formed with $\beta_3 \approx 0.11$ at scission. We note that the light fragments also have octupole deformation with $\beta_3 \approx 0.3$ – 0.4 and quadrupole deformation with $\beta_2 \approx 0.4$ – 0.8 . Such large quadrupole deformation of the light fragment is often found at scission in microscopic calculations (see, for example, figure 4 of ref. ³⁸).



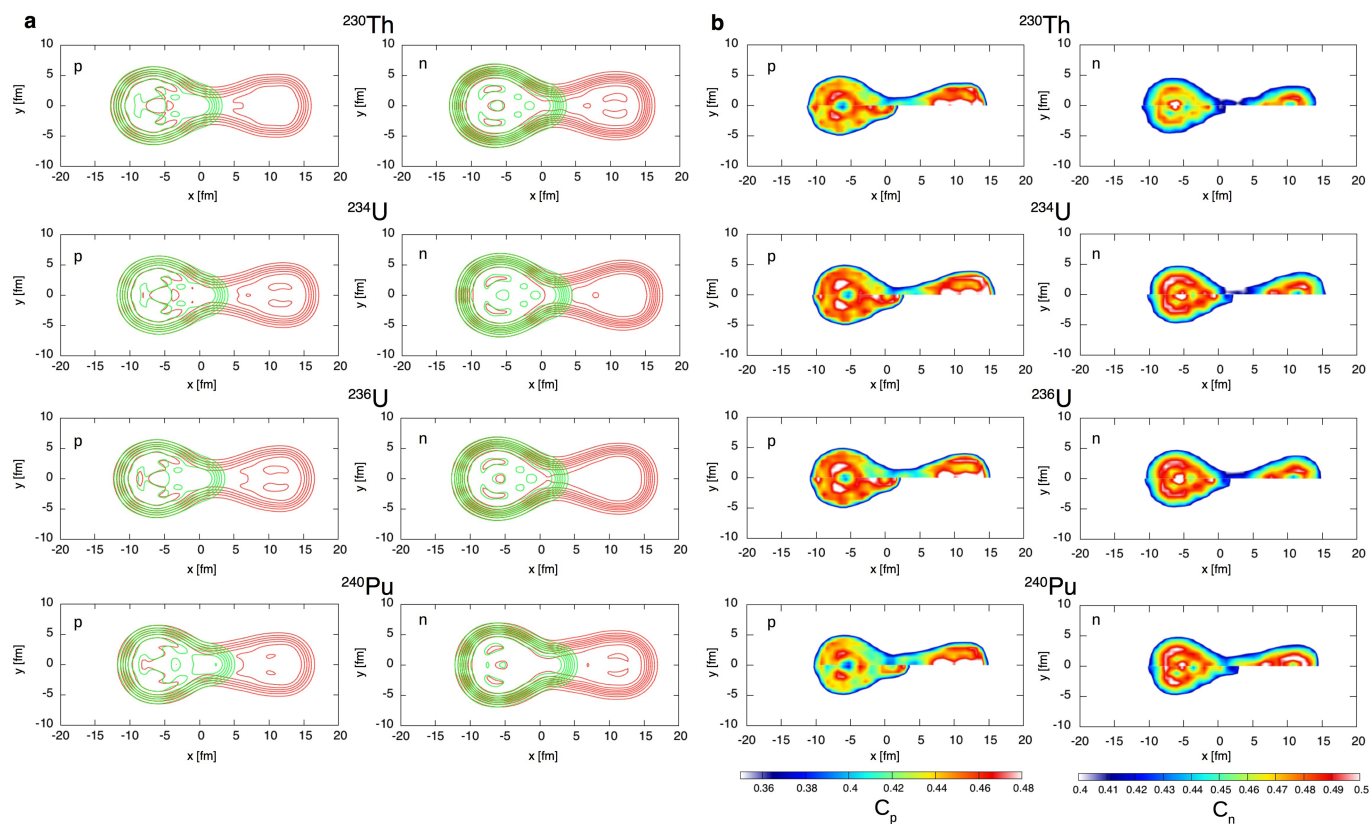
Extended Data Fig. 2 | Octupole deformation after scission. Octupole moment (see Methods) in the heavy fragment as a function of time, with a time reference ($t = 0$) corresponding to the time at which scission occurs in the calculations. In asymmetric fission of ^{258}Fm , the heavy fragment (with $Z \approx 55$) starts with a strong octupole deformation (corresponding to deformation parameter $\beta_3 \approx 0.25$ at $t = 0$) and remains octupole-deformed, possibly with different orientations (blue dashed and green dotted lines). The fragment with $Z \approx 52$ resulting from ^{246}Cm fission

(black solid line) also exhibits a substantial, yet smaller, deformation ($\beta_3 \approx 0.19$ at $t = 0$). By contrast, symmetric fission of ^{258}Fm produces Sn fragments with a much smaller octupole moment (corresponding to $\beta_3 \approx 0.11$ at $t = 0$) that oscillates around $Q_{30} = 0$ (red solid line). These results are compatible with the calculated octupole deformation energy plotted in Fig. 3b, which shows that $^{138,140}\text{Xe}$ ($Z = 54$) and ^{144}Ba ($Z = 56$) are less resistant to octupole deformation than ^{134}Te ($Z = 52$) and ^{132}Sn ($Z = 50$).



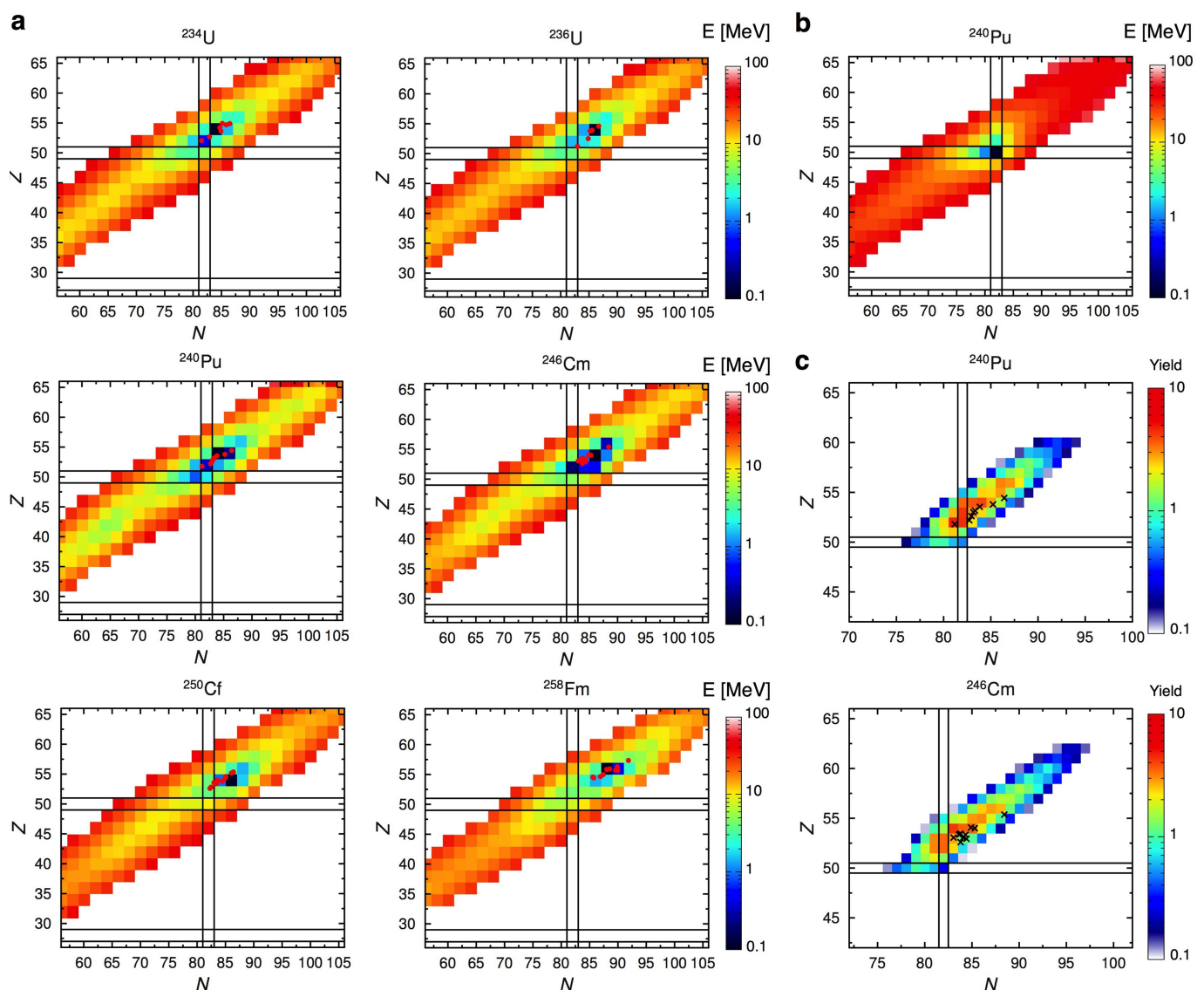
Extended Data Fig. 3 | Identification of the heavy pre-fragment in asymmetric fission of ^{258}Fm . **a**, The heavy pre-fragment is identified from its density contour using the technique of ref. ³⁹ without the assumption of reflection symmetry in the pre-fragment. Proton (left column) and neutron (right column) densities are shown with a difference of 0.01 fm^{-3} between contour lines. The fissioning asymmetric system ^{258}Fm (red lines, corresponding to calculation 8 in Extended Data Table 1) is found to form a ^{144}Ba pre-fragment with a strong octupole deformation (green lines, obtained from CHF+BCS; see Methods). **b**, Confirmation of the identification of the pre-fragment using the technique of refs ^{37,38} with a more general (that is, without assuming reflection symmetry in the pre-fragment) comparison of the proton (left column) and neutron (right

column) localization functions of ^{258}Fm (top half of each panel) and of the octupole-constrained ^{144}Ba (bottom half). The use of the deformation of ^{144}Ba as a constraint is chosen to reproduce the nucleon localization function close to the centre of the heavy fragment. The resulting octupole deformations of the ^{144}Ba pre-fragment at times $t = 0, 1.875, 3.75$ and 4.65 zs (scission occurs at 7.3 zs) are $\beta_3 \approx 0.14, 0.39, 0.39$ and 0.42 , respectively. Such strong octupole deformations could not be reached in the doubly magic ^{132}Sn nucleus without a high deformation-energy cost (25 MeV for $\beta_3 \approx 0.39$), thus hindering the formation of this fragment. The fact that the densities and localization functions of deformed ^{144}Ba match the heavy pre-fragment so well provides a clear signature of the influence of this pre-fragment before and at scission.



Extended Data Fig. 4 | Identification of the heavy pre-fragment in asymmetric fission of actinides. a, b, Same as Extended Data Fig. 3, at configurations around scission for asymmetric fission of ^{230}Th , ^{234}U , ^{236}U and ^{240}Pu . In all four systems, the heavy fragment is identified as ^{144}Ba with a constrained octupole deformation corresponding to $\beta_3 \approx 0.28$,

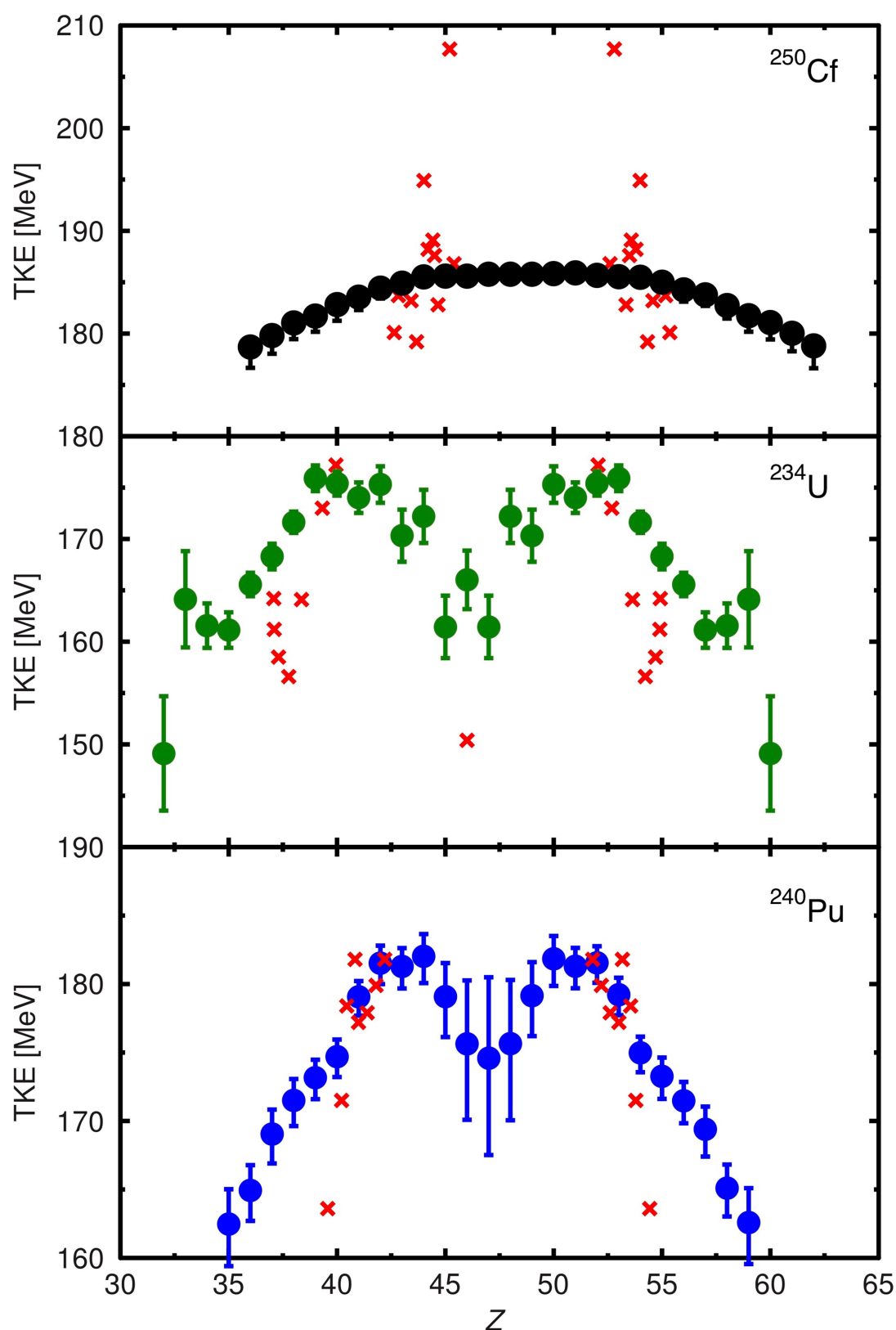
0.28, 0.27 and 0.44, respectively. The matching between deformed ^{144}Ba densities and localization functions with the heavy pre-fragment confirms the strong influence of octupole shell effects associated with $Z = 56$ and $N = 88$ on asymmetric fission.



Extended Data Fig. 5 | Effect of octupole deformation of the heavy pre-fragment on total energy at scission. **a**, To understand why the formation of a fragment is energetically more favourable in the ^{144}Ba region than in the ^{132}Sn region, we calculated the total energy of the system using a simple scission-point model⁴⁰ for various mass and charge repartitions between the fragments, each system being characterized by the number of protons Z and neutrons N in one fragment, and with the typical deformations of the fragments observed (in our TDBC calculations; see Methods) at scission. For simplicity, we only constrain the octupole deformation of the heavy fragment to be $\beta_3 = 0.35$ and the quadrupole deformation of the light fragment to be $\beta_2 = 0.6\text{--}0.8$. The binding energy of each deformed fragment is then computed from CHF+BCS simulations (see Methods) and added to the Coulomb energy between the fragments, which is approximated by the point-like formula $e^2 Z_1 Z_2 / D$ with $D = 17$ fm, where Z_1, Z_2 are the atomic numbers of the fragments, D is their distance and e is the electron charge. (As we are only interested in comparisons between different mass and charge repartitions, the strong nuclear interaction energy between the fragments is neglected because it is not expected to vary much.) The total energy $E(N, Z)$ is then plotted with its minimum value as the reference energy for each system. We note

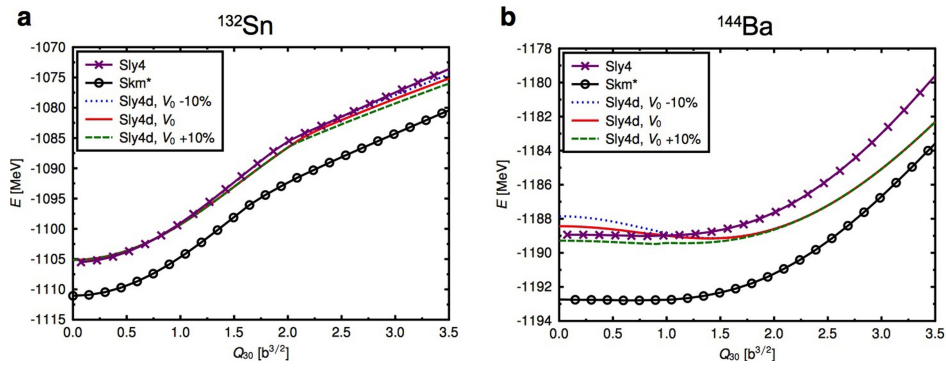
that this is a simple model that does not account for finite-temperature effects, which could potentially dampen shell effects. However, damping of shell effects is expected to occur at higher excitation energies than those involved here. Despite the simplicity of this model, the Z and N values of the fragments obtained from the TDBC calculations, shown by red dots, are clearly distributed around the system with minimum energy.

b, Same as **a**, but without the constraint on the octupole deformation of the heavy fragment (only the quadrupole deformation of the light fragment is constrained). In this case, the formation of ^{132}Sn is energetically favoured. This shows that the octupole deformation of the heavy fragment induced in the fission process strongly hinders the impact of spherical shell effects at scission. **c**, Experimental ^{240}Pu and ^{246}Cm independent fission yield (number of fragments produced after emission of prompt neutrons, but excluding radioactive decay per 100 fission reactions) from ref.³⁰ compared to the mean Z and N values obtained from TDBC calculations (black crosses). These figures show that taking into account the octupole deformation energy leads to a preference for the fragments to be formed with $Z_{\text{heavy}} \approx 54$ and overcome the effect of the spherical, doubly magic ^{132}Sn .



Extended Data Fig. 6 | Total kinetic energy of the fission fragments. TKE values obtained from TDBCS calculations (red crosses) are compared with average TKEs from experimental data^{29,41} (dots) for ^{240}Pu , ^{250}Cf and ^{234}U . As expected from the complexity of many-body dynamics, the results exhibit strong fluctuations (typically a variation of 15–20 MeV between the lowest and highest TKE for each nucleus; that is, of the same order as the experimental fluctuations of TKE). Nevertheless, the TKE values

predicted by our TDBCS calculations are essentially distributed around the average experimental TKE, indicating very good agreement between theory and experiment. For consistency, we have calculated the TKE value of a symmetric fission mode of ^{234}U (lowest red cross at $Z = 46$ in the middle panel). This calculation describes qualitatively the decrease of TKE for symmetric fission.



Extended Data Fig. 7 | Effect of functional and pairing interaction on octupole deformation. **a, b**, Deformation energy for ^{132}Sn (**a**) and ^{144}Ba (**b**) with different functionals and with pairing interaction strength V_0 (see Methods) varying by $\pm 10\%$. The Sly4 and Skm* functionals with the centre-of-mass correction and the Sly4d functional without the centre-of-mass correction give similar deformation energy curves. The pairing

interaction can slightly change the octupole deformation of the ground state of ^{144}Ba . The Sly4d functional with the normal pairing interaction (that is, with the pairing interaction strengths defined in Methods) predicts a ground-state octupole deformation of $\beta_3 = 0.165$, which is very close to the experimental value⁴ $\beta_3 = 0.17^{+0.04}_{-0.06}$.

Extended Data Table 1 | Results of TDBCS calculations

| Nucl. | # | Q_{20} [b] | E_0 [MeV] | T [zs] | $\langle Z_H \rangle$ | $\langle N_H \rangle$ | TKE | Nucl. | # | Q_{20} [b] | E_0 [MeV] | T [zs] | $\langle Z_H \rangle$ | $\langle N_H \rangle$ | TKE |
|-------------------|----|--------------|-------------|----------|-----------------------|-----------------------|-------|-------------------|----|--------------|-------------|----------|-----------------------|-----------------------|-------|
| ^{230}Th | 1 | 34.7 | 4.04 | 50.4 | 53.46 | 84.26 | 159.2 | ^{234}U | 1 | 41.0 | 1.94 | 19.9 | 52.05 | 81.5 | 177.2 |
| | 2 | 37.8 | 3.16 | 55.3 | 51.91 | 82.13 | 170.0 | | 2 | 44.2 | 0.89 | 30.8 | 54.92 | 85.16 | 164.2 |
| | 3 | 41.0 | 1.78 | 23.0 | 53.49 | 85.05 | 157.0 | | 3 | 47.3 | 0.39 | 13.6 | 54.23 | 84.7 | 156.6 |
| | 4 | 50.5 | -0.53 | 13.2 | 54.18 | 86.33 | 154.7 | | 4 | 50.5 | -0.39 | 32.4 | 54.7 | 85.86 | 158.5 |
| | 5 | 53.6 | -1.08 | 10.4 | 52.26 | 83.19 | 164.8 | | 5 | 56.8 | -2.55 | 28.0 | 52.68 | 82.9 | 173.0 |
| | 6 | 56.8 | -1.72 | 7.9 | 53.8 | 86.19 | 155.2 | | 6 | 59.9 | -3.73 | 9.5 | 54.9 | 86.47 | 161.2 |
| | 7 | 59.9 | -2.28 | 6.3 | 53.98 | 85.94 | 154.8 | | 7 | 63.1 | -4.72 | 7.6 | 53.64 | 84.9 | 164.1 |
| | 8 | 63.1 | -2.7 | 6.0 | 54.07 | 85.88 | 155.0 | ^{236}U | 1 | 41.0 | 3.03 | 27.9 | 51.2 | 82.95 | 176.6 |
| ^{246}Cm | 1 | 41.0 | 4.41 | 48.2 | 53.34 | 84.32 | 185.7 | | 2 | 47.3 | 0.28 | 30.6 | 53.9 | 85.57 | 165.0 |
| | 2 | 42.6 | 3.13 | 46.9 | 52.6 | 83.8 | 182.8 | | 3 | 50.5 | -0.53 | 32.0 | 53.78 | 85.19 | 163.6 |
| | 3 | 50.5 | -0.8 | 23.2 | 55.37 | 88.43 | 169.6 | | 4 | 56.8 | -2.5 | 13.0 | 54.6 | 86.7 | 161.3 |
| | 4 | 52.1 | -2.14 | 15.4 | 54.08 | 84.9 | 182.8 | | 5 | 59.9 | -3.5 | 13.2 | 52.5 | 84.86 | 164.5 |
| | 5 | 53.6 | -2.93 | 37.4 | 53.36 | 83.63 | 181.2 | ^{240}Pu | 1 | 45.4 | 1.46 | 20.1 | 53.79 | 85.23 | 171.5 |
| | 6 | 56.8 | -4.33 | 53.9 | 53.46 | 83.8 | 185.6 | | 2 | 46.7 | 0.8 | 16.1 | 53.17 | 83.32 | 181.8 |
| | 7 | 58.3 | -5.04 | 58.9 | 53.06 | 83.05 | 176.9 | | 3 | 50.5 | -1.16 | 89.9 | 51.8 | 81.17 | 181.8 |
| | 8 | 59.9 | -5.24 | 17.4 | 54 | 85.3 | 180.0 | | 4 | 53.0 | -2.13 | 22.5 | 52.61 | 82.95 | 177.9 |
| | 9 | 61.5 | -5.91 | 15.1 | 52.94 | 84.42 | 175.2 | | 5 | 56.8 | -3.5 | 16.0 | 53.01 | 83.15 | 177.2 |
| | 10 | 63.1 | -6.2 | 6.8 | 53.07 | 84.08 | 183.2 | | 6 | 59.3 | -4.3 | 18.0 | 53.55 | 83.83 | 178.4 |
| | 11 | 64.7 | -6.72 | 6.9 | 53.43 | 83.77 | 182.1 | | 7 | 63.1 | -5.31 | 22.1 | 54.43 | 86.42 | 163.6 |
| ^{250}Cf | 1 | 44.2 | 3.14 | 58.5 | 53.34 | 83.15 | 182.8 | | 8 | 71.9 | -7.8 | 3.4 | 52.2 | 82.72 | 179.9 |
| | 2 | 45.8 | 2.03 | 45.5 | 53.5 | 83.48 | 187.6 | ^{258}Fm | 1 | 46.3 | -1.80 | 49.3 | 54.43 | 85.72 | 190.3 |
| | 3 | 47.3 | 1.82 | 29.9 | 52.8 | 82.46 | 207.7 | | 2 | 58.7 | -6.48 | 23.2 | 54.57 | 85.58 | 186.7 |
| | 4 | 48.9 | 0.32 | 50.4 | 55.17 | 86.04 | 183.7 | | 3 | 61.9 | -8.53 | 18.3 | 57.35 | 91.86 | 180.3 |
| | 5 | 50.5 | -0.66 | 86.9 | 53.58 | 82.97 | 189.1 | | 4 | 64.9 | -10.60 | 14.1 | 55.92 | 88.54 | 182.2 |
| | 6 | 52.1 | -2.22 | 90.7 | 53.99 | 83.61 | 194.9 | | 5 | 68.0 | -12.25 | 12.4 | 56.29 | 89.90 | 178.8 |
| | 7 | 56.8 | -5.01 | 24.8 | 54.58 | 85.07 | 183.2 | | 6 | 71.2 | -13.59 | 9.2 | 55.84 | 88.00 | 181.2 |
| | 8 | 58.4 | -5.8 | 23.1 | 52.6 | 82.26 | 186.8 | | 7 | 74.3 | -14.75 | 10.7 | 54.98 | 87.42 | 183.4 |
| | 9 | 59.9 | -6.06 | 25.0 | 55.36 | 86.41 | 180.1 | | 8 | 80.5 | -16.79 | 7.3 | 54.66 | 86.93 | 182.0 |
| | 10 | 61.5 | -6.82 | 31.2 | 54.33 | 84.96 | 179.2 | | 9 | 86.7 | -18.60 | 5.7 | 55.66 | 89.97 | 177.8 |
| | 11 | 63.1 | -7.34 | 9.2 | 53.8 | 84.45 | 188.2 | | 10 | 89.8 | -19.29 | 5.6 | 55.71 | 89.61 | 176.4 |

Fissioning nucleus, simulation number, quadrupole moment Q_{20} and potential energy $E_0 = E_{\text{ini}} - E_{\text{g.s.}}$ of the fissioning system in the initial condition of the TDBCS calculation (with total energy E_{ini} and with respect to the ground-state energy $E_{\text{g.s.}}$), time T to reach scission, average proton and neutron numbers $\langle Z_H \rangle$ and $\langle N_H \rangle$ in the heavy fragment, and total kinetic energy (TKE; in MeV) of the fragments.

Measurement of the Casimir torque

David A. T. Somers^{1,2}, Joseph L. Garrett^{1,2}, Kevin J. Palm^{1,2} & Jeremy N. Munday^{1,2,3*}

Intermolecular forces are pervasive in nature and give rise to various phenomena including surface wetting¹, adhesive forces in biology^{2,3}, and the Casimir effect⁴, which causes two charge-neutral, metal objects in vacuum to attract each other. These interactions are the result of quantum fluctuations of electromagnetic waves and the boundary conditions imposed by the interacting materials. When the materials are optically anisotropic, different polarizations of light experience different refractive indices and a torque is expected to occur that causes the materials to rotate to a position of minimum energy^{5,6}. Although predicted more than four decades ago, the small magnitude of the Casimir torque has so far prevented direct measurements of it. Here we experimentally measure the Casimir torque between two optically anisotropic materials—a solid birefringent crystal (calcite, lithium niobite, rutile or yttrium vanadate) and a liquid crystal (5CB). We control the sign and strength of the torque, and its dependence on the rotation angle and the separation distance between the materials, through the choice of materials. The values that we measure agree with calculations, verifying the long-standing prediction that a mechanical torque

induced by quantum fluctuations can exist between two separated objects. These results open the door to using the Casimir torque as a micro- or nanoscale actuation mechanism, which would be relevant for a range of technologies, including microelectromechanical systems and liquid crystals.

Spatially separated, uncharged objects experience electromagnetic forces and torques due to quantum and thermal charge fluctuations. At small separations these phenomena are often called van der Waals effects and result from fluctuation-induced dipole–dipole interactions². At larger separations, greater than a few nanometres, the finite speed of light leads to retardation effects and an accompanying change in the distance dependence of the interaction, referred to as the Casimir (or Casimir–Lifshitz) regime^{4,7}. The van der Waals and Casimir effects both result from the same mechanism (quantum and thermal fluctuations), although historically they were derived from different physical pictures. A quantum-fluctuation-induced torque was predicted⁵ by considering materials with dielectric anisotropy (such as birefringent crystals); this short-range approximation was subsequently generalized⁶ to large separations. In addition to predictions of a torque arising

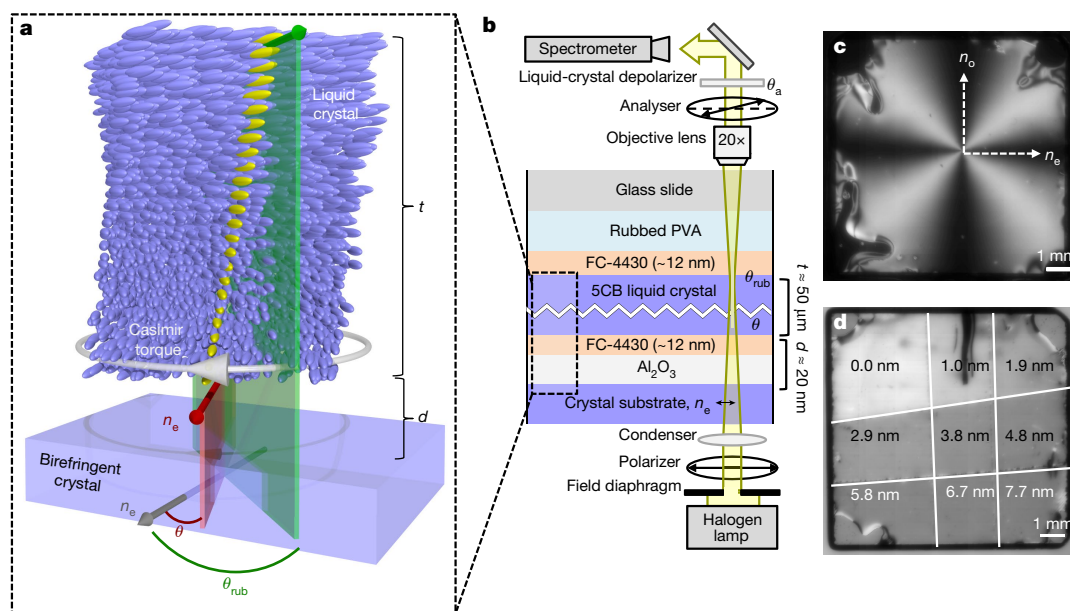


Fig. 1 | Apparatus measuring the Casimir torque. **a**, A birefringent crystal (with ordinary axis n_o and extraordinary axis n_e) causes a liquid crystal separated by a distance d to twist from θ_{rub} (anchoring at the top) to θ (final rotation angle) over a thickness $t \gg d$ (shown not to scale for clarity), owing to the Casimir torque. Yellow markers highlight the twist throughout the bulk of the liquid crystal. **b**, Schematic of the experimental set-up used to measure the torque (shown not to scale for clarity). Polarized white light from a halogen lamp is transmitted through the sample (with in-plane optical axes n_o and n_e parallel to the crystal surface) and detected via polarizing optics (containing a rotatable analyser at angle θ_a) and a spectrometer. The dashed box highlights the interacting surfaces

(liquid and solid birefringent materials separated by isotropic layers of Al_2O_3 and FC-4430). **c**, Optical micrograph of a cell with circularly rubbed PVA between crossed polarizers used to measure the dependence of the torque on θ (the angle between the principal axes of the two birefringent materials). A preferred twist of the liquid crystal towards the extraordinary axis of the substrate breaks the symmetry of the image, causing a compression of the horizontal dark regions and an expansion of the vertical dark regions. **d**, Image of a sample with linearly rubbed PVA with $\theta_{\text{rub}} = \theta_a = 45^\circ$ used to measure the dependence of the maximum torque on d at discrete, controlled separations. Lighter regions correspond to thinner Al_2O_3 layers (see labels for layer thickness) and stronger torques.

¹Department of Physics, University of Maryland, College Park, MD, USA. ²Institute for Research in Electronics and Applied Physics, University of Maryland, College Park, MD, USA. ³Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA. *e-mail: jnmunday@umd.edu

from optical anisotropy of planar objects, a torque is also predicted to exist between objects with geometric anisotropy^{8,9}. Although the Casimir force has been the subject of many theoretical and experimental investigations^{10–19}, the torque has not yet been verified experimentally. Here we report measurements of the Casimir torque between anisotropic materials separated by tens of nanometres.

The Casimir torque depends on the dielectric function of the materials, which is generally evaluated at complex frequencies $\varepsilon(i\xi)$ (ref. 6). This torque causes the two objects to align their optical axes of higher refractive index and decays as roughly d^{-2} , where d is the distance between the materials. Several experimental designs have been proposed to measure the torque, including ones involving torsion pendulums and floating microdisks^{20–24}, but no measurements have been reported. In the non-retarded regime, in which the effects of quantum fluctuations are considered to be communicated instantly between the bodies, the Casimir torque is equivalent to an anisotropic van der Waals interaction. These short-range interactions can cause alignment of nematic liquid crystals, a phase of matter with ordered molecular orientations often used in display technology. Although many commercial technologies rely on the alignment of liquid crystals on a substrate, the complex alignment mechanisms are not completely understood; however, micrometre-scale surface indentations, stretched polymer chains and optical anisotropy of the substrate are all thought to have important roles. Independent experiments^{25–28} strongly suggest that anisotropic van der Waals torques are sufficient to cause liquid-crystal anchoring. Together, these experiments demonstrated that liquid-crystal anchoring is influenced by the birefringence of a substrate and weakened by a surfactant. However, the angular and distance dependence of this interaction is unknown.

Our experimental design allows us to measure the torque via the twist of a liquid crystal (which acts as a birefringent body) in close proximity to a solid birefringent substrate with a variable spacer layer of thickness d (Fig. 1). This set-up builds on a previously proposed geometry²⁹ and enables precise optical detection of the crystal rotation³⁰. We perform measurements using a common liquid crystal (5CB) and four crystal substrates to ensure robustness of the experiments and to probe different predictions about the sign and magnitude of the Casimir torque. The substrates are calcite (CaCO_3), lithium niobate (LiNbO_3), rutile (TiO_2) and yttrium vanadate (YVO_4) coated with a thin isotropic layer of Al_2O_3 , which acts as a spacer layer (Fig. 1b, Extended Data Fig. 1). Fluorosurfactant FC-4430 (3M) is added to the 5CB at a mass concentration of 0.5% to eliminate liquid-crystal sticking at the interface²⁸. Opposite the crystal substrate is a glass slide coated in polyvinyl alcohol (PVA), which is rubbed to induce anchoring along a fixed angle θ_{rub} (all angles here are relative to the extraordinary axis of the crystal substrate; Fig. 1). The Casimir torque on the 5CB molecules near the solid birefringent substrate causes a twist that propagates through the bulk of the liquid crystal. Using the Oseen–Frank free energy³¹, the torque M per unit area A is $M/A = k_{22}\Delta\theta/t$ on the layer nearest to the crystal substrate³⁰, where $\Delta\theta = \theta_{\text{rub}} - \theta$ is the liquid-crystal twist (typically one to several degrees in our experiments), θ is the angle between the liquid-crystal director and the extraordinary axis of the crystal substrate, k_{22} is the elastic constant of the twist (3.6 ± 0.3 pN)³² and t is the measured thickness of the liquid-crystal layer (about 50 μm). For a torque of the form $M/A = a\sin(2\theta)$, the amplitude of the torque a is calculated as $a = k_{22}\Delta\theta/[t\sin(2\theta)]$.

The Casimir torque that we measure has a $\sin(2\theta)$ dependence for all four birefringent substrates, and the sign and magnitude of the torque depends on the optical properties of the crystals (Fig. 2). To probe the angular dependence, we deposit 6 nm of Al_2O_3 on four different birefringent substrates and assemble cells with circularly rubbed PVA counterplates (Methods). This produces a uniform distribution of θ_{rub} , which allows us to measure the torque as a function of θ using an optical polarimetric measurement set-up (see Methods). All four crystal substrates show a $\sin(2\theta)$ dependence of the torque, with a sign corresponding to the rotation needed to align the principal optical axes with the highest refractive index in the visible portion of the spectrum.

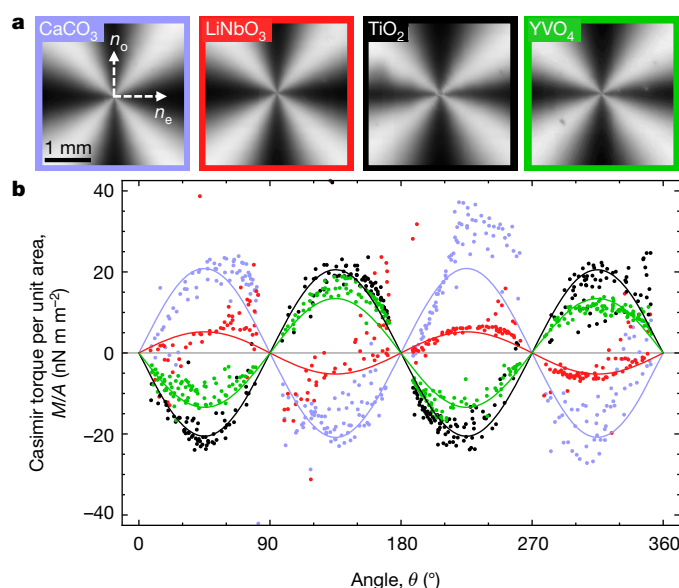


Fig. 2 | Measured $\sin(2\theta)$ angular dependence of the Casimir torque.

a, Polarized optical micrographs of cells with circularly rubbed PVA and four different birefringent substrates separated by about 18 nm from the birefringent liquid crystal. In each case, the dark brushes along the preferred axis are narrowed: ordinary axis n_o for CaCO_3 and LiNbO_3 , and extraordinary axis n_e for TiO_2 and YVO_4 . **b**, Measured torque (dots) across each cell (colour-coded) as a function of θ . Fits (solid lines) to $\sin(2\theta)$ are overlaid.

The microscope images of the samples between crossed polarizers (Fig. 2a) show the broken symmetry caused by the Casimir torque: in each case, the dark regions are narrowed along the preferred alignment axis. The data are collected at all positions, sorted into 1° bins and fitted with a $\sin(2\theta)$ function to determine the amplitude of the torque (Fig. 2b, solid lines). Control experiments with isotropic glass substrates show no evidence of a torque (Extended Data Fig. 2).

We determine the distance dependence of the Casimir torque using 27 different Al_2O_3 thicknesses (0–25 nm) per crystal substrate and measure the rotation angle to determine the maximum torque at each separation. For these experiments, the PVA layer is rubbed to $\theta_{\text{rub}} \approx 45^\circ$ and nine separations are measured on a single 1-cm² sample (Fig. 1d). Three samples are constructed for each of the four crystal types to achieve 27 distinct separations (Fig. 3). We determine the maximum torque per area for each separation to be $M_{\text{max}}/A = a\sin(90^\circ) = k_{22}\Delta\theta/[t\sin(2\theta)]$. To compare these measurements with the full Casimir-torque calculation, we include a 12-nm offset in the distance between the interacting crystals due to the surfactant, which we assume forms an isotropic layer of constant effective thickness on the substrate, and to sample roughness (see Extended Data Fig. 3 for further details). The torque is calculated using a previously described method³⁰. We find agreement between the measured (Fig. 3, symbols) and calculated (solid lines) values of the torque to within the uncertainties in the measurements and in the tabulated optical properties.

The sign of the Casimir torque depends on the optical properties of the interacting birefringent crystals and can be changed from positive to negative. We categorize these birefringent materials by the sign of the difference in refractive index $\Delta n = n_e - n_o$ and dielectric constant $\Delta\varepsilon_0 = \varepsilon_{0,e} - \varepsilon_{0,o}$ between the ordinary ('o' subscript) and extraordinary ('e' subscript) axes for wavelengths in the visible range and at zero frequency (see Extended Data Fig. 4 for optical properties). TiO_2 and YVO_4 crystals, for which $\Delta n > 0$ and $\Delta\varepsilon_0 > 0$, cause 5CB ($\Delta n > 0$, $\Delta\varepsilon_0 > 0$) to twist towards the extraordinary axis, resulting in a negative torque (Fig. 3). LiNbO_3 has $\Delta n < 0$ and $\Delta\varepsilon_0 < 0$ but weaker anisotropy, which causes a smaller twist towards the ordinary axis and a positive torque. CaCO_3 ($\Delta n < 0$, $\Delta\varepsilon_0 > 0$) is a special case: low-frequency fluctuations should contribute to a torque on 5CB towards

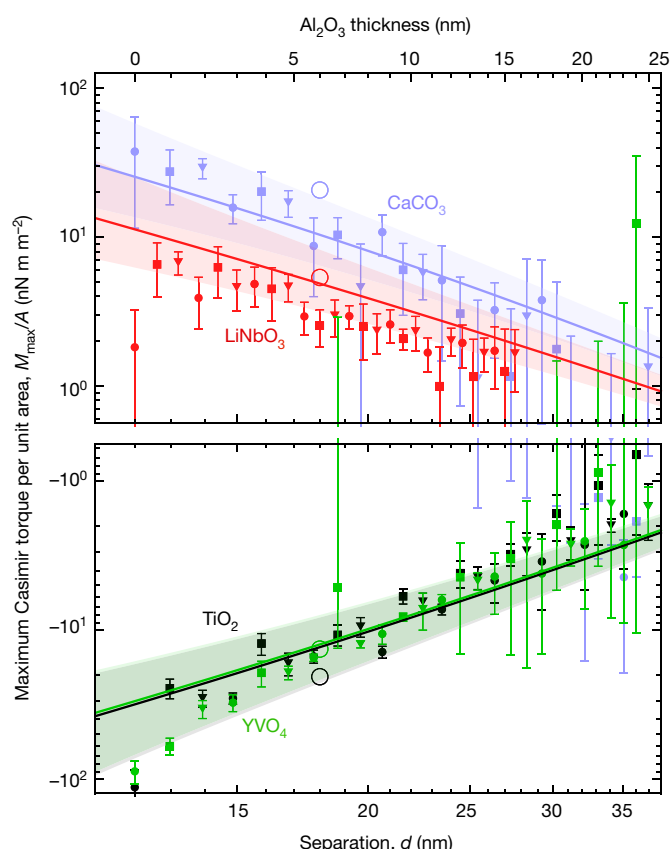


Fig. 3 | Distance dependence of Casimir torque. We measured the amplitude of the Casimir torque between 5CB and four birefringent substrates— CaCO_3 (purple), LiNbO_3 (red), TiO_2 (black) and YVO_4 (green)—as function of separation d . The separation includes the known Al_2O_3 thickness and a constant offset of 12 nm due to the surfactant and the surface roughness. Solid lines represent the calculated torque; shaded regions correspond to the range of values resulting from a range of constant offsets from 8 nm to 16 nm. Error bars denote the standard deviation of the torques measured at different locations within each region and the different types of filled symbol represent different samples of the same crystal type. The open circles represent the amplitudes of the fitted $\sin(2\theta)$ curves from Fig. 2b.

the extraordinary axis, and higher-frequency fluctuations should contribute to a torque of opposite sign towards the ordinary axis. The full Casimir-torque calculation using the available dielectric data predicts that the higher-frequency terms should dominate at the separations in our experiment, leading to a positive torque. This is consistent with our measurement, which demonstrates a torque towards the ordinary axis. At larger separations, the lower-frequency terms dominate and the sign of the torque should be reversed³³; however, no crossover behaviour is observed (or predicted) within one standard deviation of the data in the distance regime probed in our experiments. Our results are all consistent with the expected signs and relative strengths of the calculated torque. For comparison, the fitted torque amplitudes from the circularly rubbed samples (Fig. 2) are also plotted in Fig. 3 (open circles), demonstrating consistency between the different methods.

We have experimentally verified the existence of the Casimir torque between two optically anisotropic materials and have quantified the distance and angular dependence of this phenomenon. With our technique, we are able to measure torques as small as a few nN m^{-2} and have found the results to agree with calculations of the Casimir torque in terms of both sign and magnitude. The measurements presented here will permit further exploration of other theoretical predictions, such as the role of retardation and dielectric spacer layers in enhancing the torque. Finally, we anticipate that this work will open up new avenues

for using the Casimir torque in nano- and micromechanical systems, microfluidics and the self-assembly of nanostructures, in which the role of quantum vacuum fluctuations are often underexplored.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0777-8>

Received: 9 July 2018; Accepted: 25 October 2018;

Published online 19 December 2018.

- Israelachvili, J. N. *Intermolecular and Surface Forces* (Academic Press, Cambridge, 1992).
- Parsegian, V. A. *Van der Waals Forces* (Cambridge Univ. Press, Cambridge, 2005).
- Autumn, K. et al. Adhesive force of a single gecko foot-hair. *Nature* **405**, 681–685 (2000).
- Casimir, H. B. G. On the attraction between two perfectly conducting plates. *Proc. K. Ned. Akad. Wet.* **51**, 793–795 (1948).
- Parsegian, V. A. & Weiss, G. H. Dielectric anisotropy and the van der Waals interaction between bulk media. *J. Adhes.* **3**, 259–267 (1972).
- Barash, Y. S. Moment of van der Waals forces between anisotropic bodies. *Radiophys. Quantum Electron.* **21**, 1138–1143 (1978).
- Lifshitz, E. M. The theory of molecular attractive forces between solids. *J. Exp. Theor. Phys.* **2**, 73–83 (1956).
- Parsegian, V. A. Nonretarded van der Waals Interaction between anisotropic thin rods at all angles. *J. Chem. Phys.* **56**, 4393–4396 (1972).
- Hopkins, J. C., Podgornik, R., Ching, W.-Y., French, R. H. & Parsegian, V. A. Disentangling the effects of shape and dielectric response in van der Waals interactions between anisotropic bodies. *J. Phys. Chem. C* **119**, 19083–19094 (2015).
- Lamoreaux, S. K. Demonstration of the Casimir force in the 0.6 to 6 μm range. *Phys. Rev. Lett.* **78**, 5–8 (1997).
- Munday, J. N., Capasso, F. & Parsegian, V. A. Measured long-range repulsive Casimir–Lifshitz forces. *Nature* **457**, 170–173 (2009).
- Woods, L. M. et al. Materials perspective on Casimir and van der Waals interactions. *Rev. Mod. Phys.* **88**, 045003 (2016).
- Chan, H. B. et al. Measurement of the Casimir force between a gold sphere and a silicon surface with nanoscale trench arrays. *Phys. Rev. Lett.* **101**, 030401 (2008).
- Banisher, A. A., Wagner, J., Emig, T., Zandi, R. & Mohideen, U. Demonstration of angle-dependent Casimir force between corrugations. *Phys. Rev. Lett.* **110**, 250403 (2013).
- Bordag, M., Klimchitskaya, G. L., Mohideen, U. & Mostepanenko, V. M. *Advances in the Casimir Effect* (Oxford Univ. Press, Oxford, 2009).
- Milton, K. A. *The Casimir Effect* (World Scientific, New York, 2001).
- Milonni, P. W. *The Quantum Vacuum: An Introduction to Quantum Electrodynamics* (Academic Press, Cambridge, 1993).
- Decca, R. S., López, D., Fischbach, E. & Krause, D. E. Measurement of the Casimir force between dissimilar metals. *Phys. Rev. Lett.* **91**, 050402 (2003).
- Torricelli, G. et al. Switching Casimir forces with phase-change materials. *Phys. Rev. A* **82**, 010101 (2010).
- Chen, X. & Spence, J. C. H. On the measurement of the Casimir torque. *Phys. Scripta* **248**, 2064–2071 (2011).
- Guérout, R., Genet, C., Lambrecht, A. & Reynaud, S. Casimir torque between nanostructured plates. *Europhys. Lett.* **111**, 44001 (2015).
- Munday, J. N., Iannuzzi, D., Barash, Y. & Capasso, F. Torque on birefringent plates induced by quantum fluctuations. *Phys. Rev. A* **71**, 042102 (2005).
- Munday, J. N., Iannuzzi, D. & Capasso, F. Quantum electrodynamical torques in the presence of Brownian motion. *New J. Phys.* **8**, 244 (2006).
- Rodrigues, R. B., Neto, P. A. M., Lambrecht, A. & Reynaud, S. Vacuum-induced torque between corrugated metallic plates. *Europhys. Lett.* **76**, 822–828 (2006).
- Dubois-Violette, E. & De Gennes, P. G. Effects of long range van der Waals forces on the anchoring of a nematic fluid at an interface. *J. Colloid Interface Sci.* **57**, 403–410 (1976).
- Schadt, M., Schmitt, K., Kozinkov, V. & Chigrinov, V. Surface-induced parallel alignment of liquid crystals by linearly polymerized photopolymers. *Jpn. J. Appl. Phys.* **31**, 2155–2164 (1992).
- Lu, M. Liquid crystal orientation induced by van der Waals interaction. *Jpn. J. Appl. Phys.* **43**, 8156–8160 (2004).
- Bryan-Brown, G. P., Wood, E. L. & Sage, I. C. Weak surface anchoring of liquid crystals. *Nature* **399**, 338–340 (1999).
- Smith, E. R. & Ninham, B. W. Response of nematic liquid crystals to van der Waals forces. *Physica* **66**, 111–130 (1973).
- Samers, D. A. T. & Munday, J. N. Rotation of a liquid crystal by the Casimir torque. *Phys. Rev. A* **91**, 032520 (2015).
- Frank, F. Liquid crystals. On the theory of liquid crystals. *Discuss. Faraday Soc.* **25**, 19–28 (1958).
- Toyooka, T., Chen, G., Takezoe, H. & Fukuda, A. Determination of twist elastic constant K_{22} in 5CB by four independent light-scattering techniques. *Jpn. J. Appl. Phys.* **26**, 1959–1966 (1987).
- Thiyam, P. et al. Distance-dependent sign reversal in the Casimir–Lifshitz torque. *Phys. Rev. Lett.* **120**, 131601 (2018).

Acknowledgements This work was supported by the National Science Foundation under grant numbers PHY-1506047 and PHY-1806768. We acknowledge support from the FabLab at the Maryland NanoCenter and thank M. S. Leite for discussions and comments on the manuscript.

Reviewer information *Nature* thanks O. Lavrentovich, S. Zumer and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.N.M. conceived and supervised the project. D.A.T.S. designed the apparatus, performed experiments and analysed the resultant data. J.L.G. and K.J.P. performed AFM experiments and analysis, and K.J.P. performed spectroscopic ellipsometric measurements and analysis. All authors discussed and interpreted the data and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0777-8>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0777-8>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to J.N.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Atomic layer deposition. The Al_2O_3 layers are deposited by atomic layer deposition (Beneq) at 150 °C and 3 mbar. The growth rate is measured by deposition on silicon wafers to be 0.12 nm per cycle, and the thickness varies by less than 1 nm over a 1-cm² region. To change the Al_2O_3 thickness on a single sample, different regions are masked using Kapton tape. For each type of crystal, three substrates are coated identically with nine different thicknesses (in intervals of 3j cycles, where $j = 5$ for LiNbO_3 and $j = 8$ for the other crystals) in four deposition steps (Fig. 1d). With two more deposition steps, the three substrates receive an additional deposited layer of 0, j or 2j cycles, resulting in interlaced Al_2O_3 thicknesses across the three substrates. Before each deposition step, the substrates are cleaned with acetone, methanol and isopropanol, dried with N_2 , and placed in a UV/ozone cleaner (Ossila) for 3 min to remove any remaining contaminants.

Construction of the liquid-crystal cell. To make the rubbed PVA slides, borosilicate glass cover slips are cleaned using the same procedure described above. They are spin-coated with 1% PVA in water at 2,000 r.p.m. for 30 s and dried on a hot plate at 105 °C for 30 min. They are then rubbed with a velvet cloth to cause strong liquid-crystal anchoring at the interface (micrometre-scale surface indentations, stretched polymer chains and optical anisotropy of the substrate are all thought to have important roles in the alignment^{34,35}). For circular cells (Figs. 1c and 2), the glass slides are mounted onto a motor that spins them while the cloth is lowered to contact for around 30 revolutions. For linearly rubbed cells (Figs. 1d and 3), the slides are rubbed 30 times unidirectionally. The cells are assembled by cleaning the Al_2O_3 -coated crystals with solvents and UV/ozone as above, placing dots of UV-curable glue (Norland Optics NOA68) mixed with 50- μm spacer beads (Cospheric) in the corners of the sample. They are then pressed to the rubbed PVA slide, cured for 5 min with UV light, placed onto a hotplate at 55 °C, and filled with the liquid-crystal mixture in the isotropic phase (0.5% FC-4430 in 5CB) via capillary action. The hot plate is then turned off and the samples are allowed to cool to room temperature for 1 h before measurement. The FC-4430 covers both surfaces and probably reduces the anchoring strength at the rubbed PVA layer as well as at the crystalline substrate. However, the anchoring at the rubbed PVA remains many orders of magnitude stronger than the Casimir torque, and the liquid-crystal alignment direction remains indistinguishable from the rubbing direction at the top of the cell.

Optical polarimetric measurement. We measure the liquid-crystal anchoring with polarized microscopy (Nikon Eclipse Ti-U; Fig. 1b). First, the sample is rotated so that the extraordinary axis of the solid crystal is aligned with a fixed polarizer. With this orientation, the polarized light passes through the solid crystal without any effect from its birefringence. The illuminated spot is narrowed to a 400- μm diameter with an aperture iris to permit spatial mapping of the liquid-crystal anchoring. An analyser is then rotated in 7.5° increments and the transmitted light passes through a custom-made liquid-crystal depolarizer before being directed into a spectrometer (Thorlabs CCS175). The intensity of the transmitted light as a function of wavelength and analyser angle (θ_a) is recorded (Extended Data Fig. 5b, c) and modelled using Jones calculus (Extended Data Fig. 5d, e):

$$I = \cos^2(X) + \cos^2(\Delta\theta - \theta_a) + \Delta\theta \cos(X) \sin[2(\Delta\theta - \theta_a)] \frac{\sin(X)}{X} + \left[(\Delta\theta)^2 \sin^2(\Delta\theta - \theta_a) + \frac{(\Delta\varphi)^2}{4} \cos^2(\Delta\theta - \theta_a + 2\theta) \right] \frac{\sin^2(X)}{X^2} \quad (1)$$

where $X = \sqrt{\Delta\varphi^2/4 + \Delta\theta^2}$ and $\Delta\varphi = 2\pi\Delta n \times t/\lambda$ is the optical retardance of the liquid-crystal layer with birefringence Δn , thickness t and wavelength λ . A nonlinear fitting algorithm is used to extract $\Delta\theta$ and θ at each point. To obtain a fit, t is first fitted using the known birefringence of 5CB³⁶, and then the birefringence Δn is allowed to vary slightly (at most 0.3%) with fixed t in the nonlinear fit. This step is necessary because small errors in Δn (for example, due to small temperature variations) make a reasonable fit impossible. As a consistency check, fits with Δn values that vary by more than 0.3% at any wavelength from the known curve are rejected.

Data analysis. Using a motorized microscope stage, about 625 local polarized spectrometry measurements were taken across each 1-cm² cell (Extended Data Fig. 5).

We select the data that correspond to a particular region (such as that corresponding to a single Al_2O_3 thickness). Extreme outliers (those that lie more than six times the interquartile range from the median), which can arise from local defects in the cell, are rejected. The remaining data are summarized with a mean and standard deviation.

Measurement sensitivity. In order of decreasing importance, the sensitivity of our measurement is limited by non-uniformities in the liquid-crystal cells, the optical measurement technique and Brownian motion. The first of these is the dominating factor in our measurements and can be estimated by the large-separation torques in Fig. 3. At large separations for LiNbO_3 , TiO_2 and YVO_4 , we can resolve torques with magnitudes of about $3 \times 10^{-9} \text{ N m m}^{-2}$ (several orders of magnitude smaller than typical ultraweak liquid-crystal anchoring energies; the FC-4430 is used to greatly reduce these energies so that the Casimir-torque energy is the dominant interaction). The CaCO_3 substrates are more susceptible to surface defects in fabrication, which limits the sample quality.

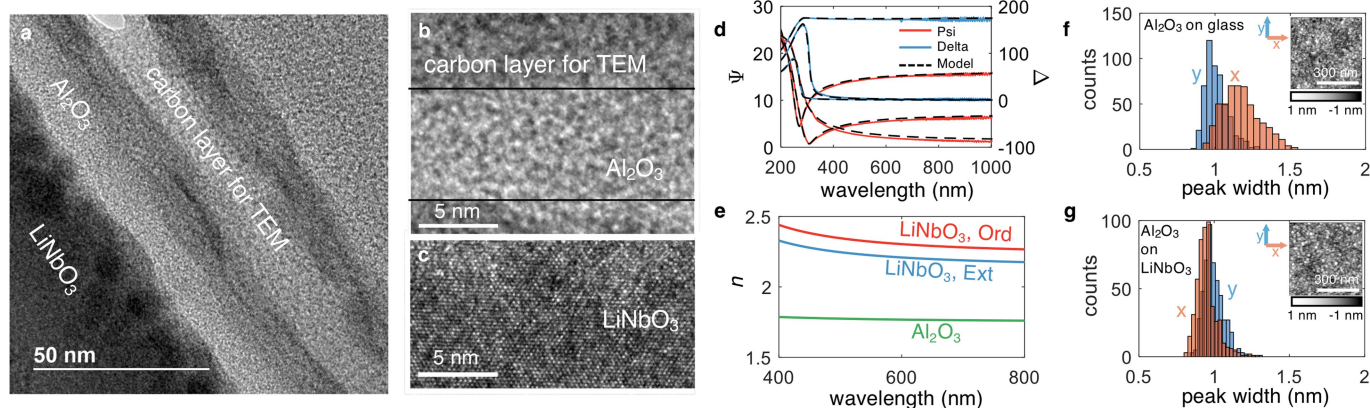
With an ideal fabrication process that produces uniform and defect-free samples, the optical measurement would limit the sensitivity. From the torque balance equation $a \sin(2\theta) = k_{22}\Delta\theta/t$, the smallest measurable torque is limited by the smallest measurable twist angle $\Delta\theta$. For a cell with $t = 50 \mu\text{m}$, a measurement of $\Delta\theta$ with a resolution of 0.1° corresponds to a torque sensitivity of around $10^{-10} \text{ N m m}^{-2}$. The resolution of the torque measurement can be improved by increasing t , but doing so introduces other practical experimental issues, such as polarization-dependent absorption and long-range liquid-crystal defects. For very small values of t (less than 1 μm), the torque resolution would become worse and other fluctuation-induced effects, such as the critical Casimir effect, would need to be considered if the temperature of the cell approaches the liquid-crystal phase transition³⁷.

If all other factors were mitigated, the minimum noise level of our measurement technique would be dominated by thermal noise of magnitude $k_B T$. The measured torque per unit area is $k_{22}\Delta\theta/t$ and the free energy associated with the dependent variable $\Delta\theta$ is $k_{22}(\Delta\theta)^2/(2t)$. Setting this equal to $k_B T/A$ (where A is the area of the measured spot size), we find that the minimum resolvable $\Delta\theta$ due to thermal noise is $\sqrt{k_B T/(Ak_{22})}$. With $T = 300 \text{ K}$, $t = 50 \mu\text{m}$ and $A = 1 \text{ mm}^2$, this is about 0.01°, which corresponds to a best possible torque-per-unit-area sensitivity of about $10^{-11} \text{ N m m}^{-2}$.

Data availability

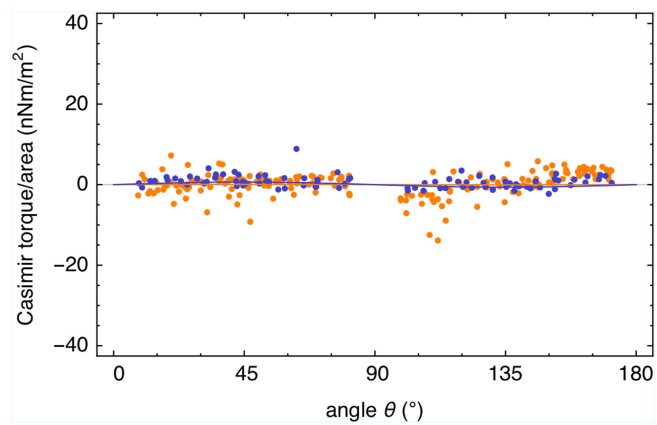
The data that support the findings of this study are available from the corresponding author on reasonable request.

- Ishihara, S. How far has the molecular alignment of liquid crystals been elucidated? *J. Disp. Technol.* **1**, 30–40 (2005).
- Kobayashi, S., Kuroda, K., Matsuo, M. & Nishikawa, M. In *The Liquid Crystal Display Story* (ed. Koide, N.) 59–80 (Springer, Tokyo, 2014).
- Li, J. & Wu, S. T. Extended Cauchy equations for the refractive indices of liquid crystals. *J. Appl. Phys.* **95**, 896–901 (2004).
- Li, H. & Kardar, M. Fluctuation-induced forces between rough surfaces. *Phys. Rev. Lett.* **67**, 3275–3278 (1991).
- Miikkulainen, V., Leskelä, M., Ritala, M. & Puurunen, R. L. Crystallinity of inorganic films grown by atomic layer deposition: overview and general trends. *J. Appl. Phys.* **113**, 021301 (2013).
- Hu, J., Xiao, X. D., Ogletree, D. F. & Salmeron, M. Imaging the condensation and evaporation of molecularly thin films of water with nanometer resolution. *Science* **268**, 267–269 (1995).
- Neuman, K. C. & Nagy, A. Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nat. Methods* **5**, 491–505 (2008).
- Kornilovitch, P. E. Van der Waals interaction in uniaxial anisotropic media. *J. Phys. Condens. Matter* **25**, 035102 (2013); corrigendum **30**, 189501 (2018).
- Hough, D. B. & White, L. R. The calculation of hamaker constants from liftshitz theory with applications to wetting phenomena. *Adv. Colloid Interface Sci.* **14**, 3–41 (1980).
- Shi, H.-s., Zhang, G. & Shen, H.-y. Measurement of principal refractive indices and the thermal refractive index coefficients of yttrium vanadate. *J. Synth. Cryst.* **30**, 85–88 (2001).
- Vali, R. Ab initio vibrational and dielectric properties of. *Solid State Commun.* **149**, 1637–1640 (2009).

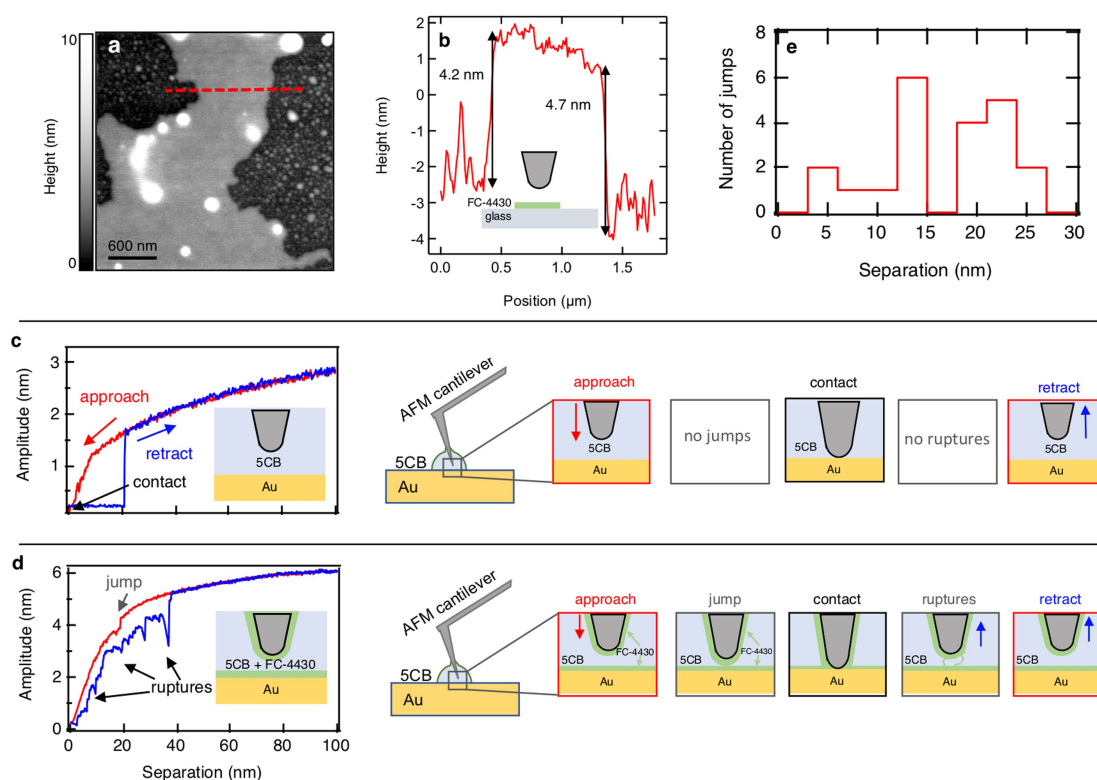


Extended Data Fig. 1 | Evidence for amorphous, isotropic deposition of Al_2O_3 . Atomic layer deposition of Al_2O_3 on planar substrates at low temperatures (in our case 150 °C) generally results in isotropic, amorphous films³⁸. To confirm this expectation, we performed three different experiments to characterize the films. **a–c**, First, TEM measurements at different magnifications show that the Al_2O_3 layer is amorphous (**a**, **b**), whereas the LiNbO_3 layer is crystalline (**a**, **c**), as expected. **d**, Second, spectroscopic ellipsometry data are found to be consistent with an isotropic layer of Al_2O_3 existing on top of the birefringent LiNbO_3 crystal. We first determined the optical properties for the birefringent LiNbO_3 crystal without any Al_2O_3 coating. Ψ and Δ data (ellipsometric amplitude and phase ratio, respectively) for 55°, 60° and 75° measurement angles are performed and fitted to a Tauc Lorentz biaxial (anisotropic) model, yielding a good ellipsometric fit. **e**, Third, spectroscopic ellipsometry was performed on a LiNbO_3 crystal with a 5-nm-thick Al_2O_3 coating. The previously determined model for LiNbO_3 was used together with an isotropic model for Al_2O_3 , yielding a similarly good fit with no increase in the mean squared error. A birefringent model for Al_2O_3 could also be

used; however, this results in additional fit parameters that artificially reduce the mean squared error below that obtained with the LiNbO_3 crystal alone, providing strong evidence that the isotropic model of Al_2O_3 is more physical. The resulting indices of refraction n are shown for both the ordinary ('Ord') and extraordinary ('Ext') axes of LiNbO_3 and for isotropic Al_2O_3 . **f**, **g**, Fourth, to rule out potential in-plane anisotropy due to correlation in the orientation of the surface roughness, we performed atomic force microscopy (AFM) topography scans of the Al_2O_3 surface as-deposited for our glass control samples (**f**) and the LiNbO_3 crystal (**g**). The roughness is randomly oriented with no preferred direction in both cases. We further considered the widths of the roughness peaks in the two perpendicular directions (x and y in the insets, which show the topography scans). For the Al_2O_3 film on the glass substrate, the average peak widths in the x and y directions are 1.2 ± 0.3 nm and 1.0 ± 0.2 nm, respectively. Similarly, for the Al_2O_3 film on the LiNbO_3 substrate, the average peak widths in the x and y directions are both 1.0 ± 0.2 nm. Although small variations between the x and y scans may be expected due to sample drift or tip geometry, no substantial difference is measured.

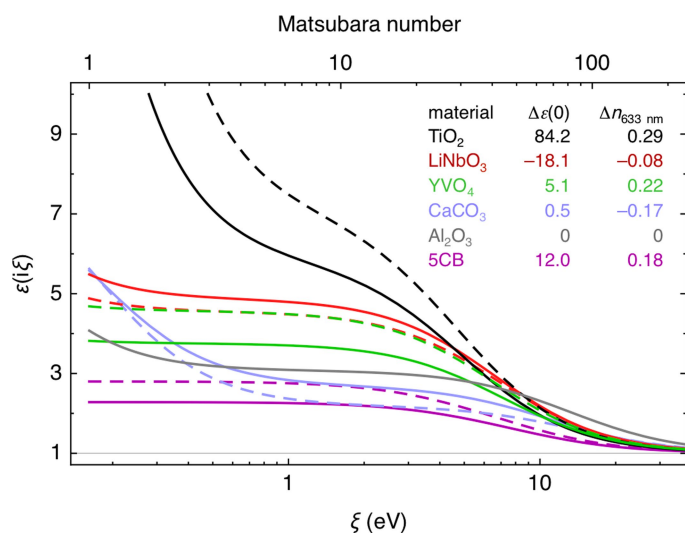


Extended Data Fig. 2 | Control experiment showing no measured torques from isotropic borosilicate glass. Glass substrates are coated with 6 nm of Al_2O_3 (orange) or around 6 nm of PVA (purple). When fitted with a $\sin(2\theta)$ function (solid lines), there is no measurable torque.



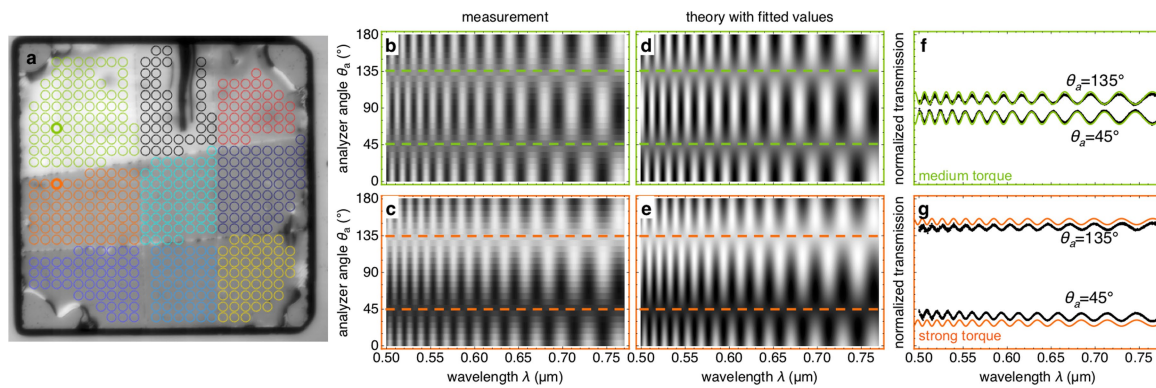
Extended Data Fig. 3 | Determination of the distance offset. We use AFM to probe the thickness of the FC-4430 surfactant layer that forms between the 5CB liquid crystal and the solid substrate, using two different methods^{39,40}. **a**, The first sample was prepared by spreading a 0.5% FC-4430 in 5CB mixture (similar to those used in the torque experiments) across a glass substrate to allow for segregation of the materials. AFM scans (dynamic mode in air to ensure that the optical lever detection scheme of AFM was not distorted by the birefringence of the liquid crystal) show the surfactant monolayer (grey) on glass (black) with 5CB droplets (white). **b**, A line scan across surfactant layer (dashed red line in **a**) shows the step height of the surfactant film on glass to be 4.5 ± 0.3 nm. This measurement puts a minimum bound on the thickness of the surfactant layer, because when immersed in 5CB (as done in the torque experiments) the surfactant molecules will probably extend further from the surface. **c**, **d**, To determine the total extent of the surfactant in situ, we performed a second set of experiments, using a technique similar to that used in ref. ⁴⁰. For these experiments, a gold sample is used to allow a bias voltage to be applied between the tip and the sample to more accurately determine the exact location of tip-sample contact. **c**, A droplet of 5CB is

placed on the gold sample and a platinum AFM tip is brought towards the surface under 3-V a.c. bias, without any surfactant present. The amplitude of the piezoelectrically driven oscillation of the tip smoothly decreases as it approaches the 5CB-gold boundary, except where the tip sticks to the surface after contact (tip-sample separation of zero). **d**, A second gold sample is used with a droplet of the 0.5% FC-4430 in 5CB mixture (similar to that used in the torque experiments), so that the surfactant covers both surfaces. Again, the amplitude of tip oscillation is recorded during the approach and retract runs under 3-V a.c. bias. On approach, the amplitude decreases rapidly at a certain separation (about 20 nm for this scan), indicating the distance at which the surfactant layers on each surface join together (that is, 'jump' to contact). On retraction, multiple features are observed at locations where molecules attached to both surfaces separate (that is, 'rupture'). **e**, Histogram showing separations at which jumps occur over several distinct distance sweeps. A jump is observed in 21 of 31 approach-retract curves and occurs at an average tip-sample separation of 16.6 nm, which would imply a 8.3-nm-thick surfactant layer on each surface when immersed in 5CB. Considering the measured surface roughness of 4–5 nm, we use a distance offset of 12 ± 4 nm.



Extended Data Fig. 4 | Dielectric models for the relevant materials.

For birefringent materials, solid and dashed lines indicate the dielectric function ϵ along the ordinary and extraordinary axes, respectively. The dielectric models for the 5CB liquid crystal are from ref. ⁴¹, the dielectric functions for the solid crystals are modelled using a previous method⁴² and the dielectric data for YVO₄ are from refs ^{43,44}. The dielectric function of the thin FC-4430 layer is unknown, but its precise value has little effect on the torque. For our calculations, the layers between the two crystals are treated as a uniform, homogeneous medium described by the optical properties of Al₂O₃ alone (rather than part Al₂O₃ and part FC-4430). As an extreme case, we also calculate the torque with a uniform medium with the optical properties of H₂O, which are probably similar to the unknown FC-4430, rather than Al₂O₃. The calculated torque changes by less than 6% for the distances used in our experiment and are thus within our calculation uncertainties.



Extended Data Fig. 5 | Polarized spectrometry measurement of each sample. **a**, Polarized white-light image of a 1-cm² substrate (TiO₂) with nine different Al₂O₃ thicknesses achieved through atomic layer deposition and masking. The coloured circles indicate separate polarized spectrometry measurements, two of which (bolded) are shown in more detail in **b–g**. **b, c**, Normalized, measured transmission as a function of analyser angle and wavelength for regions with thick (**b**) and thin (**c**) Al₂O₃ layers. The thick (thin) layer corresponds to a larger (smaller) separation d between the birefringent materials. The dashed lines indicate θ_{rub} and

$\theta_{rub} - 90^\circ$. **d, e**, Results from fitting **b** and **c** using equation (1) in Methods. **f, g**, Transmitted intensities (dots, measured values; solid lines, theoretical fits) along $\theta_a = \theta_{rub} = 135^\circ$ and $\theta_a = \theta_{rub} - 90^\circ = 45^\circ$ are shown for the thick (**f**) and thin (**g**) Al₂O₃ layers. A strong torque causes a larger difference in transmission between $\theta_a = \theta_{rub}$ and $\theta_a = \theta_{rub} - 90^\circ$ (compare **g** and **f**). The slight offset between the fit and measurement in **g** is because only two of the 25 measurements used in the fit are plotted. Combining all measurements, the error in the angle is less than 4° .

Electric-field-tuned topological phase transition in ultrathin Na₃Bi

James L. Collins^{1,2,3}, Anton Tadich^{3,4}, Weikang Wu⁵, Lidia C. Gomes^{6,7}, Joao N. B. Rodrigues^{6,8}, Chang Liu^{1,2,3}, Jack Hellerstedt^{1,2,9}, Hyejin Ryu^{10,11}, Shujie Tang¹⁰, Sung-Kwan Mo¹⁰, Shaffique Adam^{6,12}, Shengyuan A. Yang^{5,13}, Michael S. Fuhrer^{1,2,3} & Mark T. Edmonds^{1,2,3*}

The electric-field-induced quantum phase transition from topological to conventional insulator has been proposed as the basis of a topological field effect transistor^{1–4}. In this scheme, ‘on’ is the ballistic flow of charge and spin along dissipationless edges of a two-dimensional quantum spin Hall insulator^{5–9}, and ‘off’ is produced by applying an electric field that converts the exotic insulator to a conventional insulator with no conductive channels. Such a topological transistor is promising for low-energy logic circuits⁴, which would necessitate electric-field-switched materials with conventional and topological bandgaps much greater than the thermal energy at room temperature, substantially greater than proposed so far^{6–8}. Topological Dirac semimetals are promising systems in which to look for topological field-effect switching, as they lie at the boundary between conventional and topological phases^{3,10–16}. Here we use scanning tunnelling microscopy and spectroscopy and angle-resolved photoelectron spectroscopy to show that mono- and bilayer films of the topological Dirac semimetal^{3,17} Na₃Bi are two-dimensional topological insulators with bulk bandgaps greater than 300 millielectronvolts owing to quantum confinement in the absence of electric field. On application of electric field by doping with potassium or by close approach of the scanning tunnelling microscope tip, the Stark effect completely closes the bandgap and re-opens it as a conventional gap of 90 millielectronvolts. The large bandgaps in both the conventional and quantum spin Hall phases, much greater than the thermal energy at room temperature (25 millielectronvolts), suggest that ultrathin Na₃Bi is suitable for room-temperature topological transistor operation.

Two-dimensional quantum spin Hall (QSH) insulators are characterized by an insulating interior with bulk bandgap E_g , and topologically protected conducting edge channels that are robust to backscattering by non-magnetic disorder. The QSH effect was first realized in HgTe quantum wells⁵ where the small E_g prevents device applications above cryogenic temperatures. This has led to efforts to find new materials with $E_g \gg 25$ millielectronvolts (meV) for room-temperature topological electronic devices (25 meV is the thermal energy at room temperature). Recent reports of QSH insulators bismuthene on SiC ($E_g \approx 0.8$ eV)⁶ and monolayer 1T'-WTe₂ ($E_g \approx 50$ meV)⁷ are promising, with the QSH effect measured in monolayer WTe₂ up to ⁹100 K. However, a predicted electric-field effect in WTe₂ has not yet been reported experimentally, and a substantial field effect in atomically two-dimensional (2D) bismuthene is unlikely owing to the completely in-plane structure, which suggests that any Stark effect would most probably be small.

Ultrathin films of topological Dirac semimetals (TDSs) are a promising material class to realize the electric-field-tuned topological

phase transition, with such a transition predicted³ in few-layer films of TDS Na₃Bi and Cd₃As₂. Bulk TDSs are zero-bandgap semimetals with a linear band dispersion in all three dimensions around pairs of Dirac points^{10–13}, whereas few-layer TDSs are predicted¹⁷ to be non-trivial insulators with bulk bandgaps up to about 300 meV for monolayer Na₃Bi. However, experiments on few-layer TDSs are lacking at present, with only 10–15-nm thin films grown to date^{14–16}. To unambiguously demonstrate electric field control over the magnitude, electric field-dependence, and topological nature of the bandgap in ultrathin Na₃Bi, we employ two independent experimental techniques. First, we utilize angle-resolved photoelectron spectroscopy (ARPES) to measure directly the electronic band structure and its modification as a result of doping the surface with potassium (K) to generate an electric field. Second, we use scanning tunnelling spectroscopy (STS), which measures the local density of states (LDOS) as a function of energy, to probe the energy gap directly while varying the tip-sample separation and consequently the induced electric field caused by the potential difference between tip and sample. STS also resolves the topological edge state in Na₃Bi at low electric field, demonstrating the topological nature of this phase. These experimental observations are well supported by density-functional theory (DFT) band structure and edge state calculations with and without electric field.

The unit cell of Na₃Bi contains two stacked triple layers in the *z* direction, comprising Na and Bi atoms that form a honeycomb structure, with interleaved Na atoms, as shown in the crystal structures of Fig. 1a, b. One triple layer and two stacked triple layers correspond to monolayer (ML) and bilayer (BL) Na₃Bi respectively, as illustrated in Fig. 1b. The symmetry groups of pristine ML and BL Na₃Bi, and the result of including a Na surface vacancy are summarized in Methods and Extended Data Fig. 11. In Fig. 1c, scanning tunnelling microscopy (STM) on few-layer Na₃Bi(001) epitaxial films grown via molecular beam epitaxy (MBE) on Si(111) (see Methods for details) reveals coexisting regions of ML and BL Na₃Bi islands that are atomically flat and up to 40 nm in size, along with small areas of bare substrate. Monolayer regions are identified by an additional 0.22 nm distance to the underlying substrate, due to interfacial spacing or structural relaxation which has been observed previously in other atomically thin materials¹⁸. Figure 1d shows the overall band structure of few-layer Na₃Bi films along the \bar{M} – $\bar{\Gamma}$ – \bar{K} surface directions measured with ARPES at $h\nu = 48$ eV, along with the 2D Brillouin zone (Fig. 1d inset). Figure 1e shows the second derivative of the spectra in order to enhance low-intensity features. This has been overlaid with DFT calculations for ML (blue) and BL (green) Na₃Bi showing qualitatively good agreement, consistent with the STM topography which shows coexisting ML and BL regions. Photon-energy dependent ARPES (see Extended Data

¹School of Physics and Astronomy, Monash University, Clayton, Victoria, Australia. ²Monash Centre for Atomically Thin Materials, Monash University, Clayton, Victoria, Australia. ³ARC Centre of Excellence in Future Low-Energy Electronics Technologies, Monash University, Clayton, Victoria, Australia. ⁴Australian Synchrotron, Clayton, Victoria, Australia. ⁵Research Laboratory for Quantum Materials, Singapore University of Technology and Design, Singapore, Singapore. ⁶Department of Physics and Centre for Advanced 2D Materials, National University of Singapore, Singapore, Singapore. ⁷National Centre for Supercomputing Applications, University of Illinois at Urbana-Champaign, Champaign, IL, USA. ⁸Institute for Condensed Matter Theory and Department of Physics, University of Illinois at Urbana-Champaign, Champaign, IL, USA. ⁹Institute of Physics of the Czech Academy of Sciences, Prague, Czech Republic. ¹⁰Advanced Light Source, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. ¹¹Center for Spintronics, Korea Institute of Science and Technology, Seoul, South Korea. ¹²Yale-NUS College, Singapore, Singapore. ¹³Centre for Quantum Transport and Thermal Energy Science, School of Physics and Technology, Nanjing Normal University, Nanjing, China. *e-mail: mark.edmonds@monash.edu

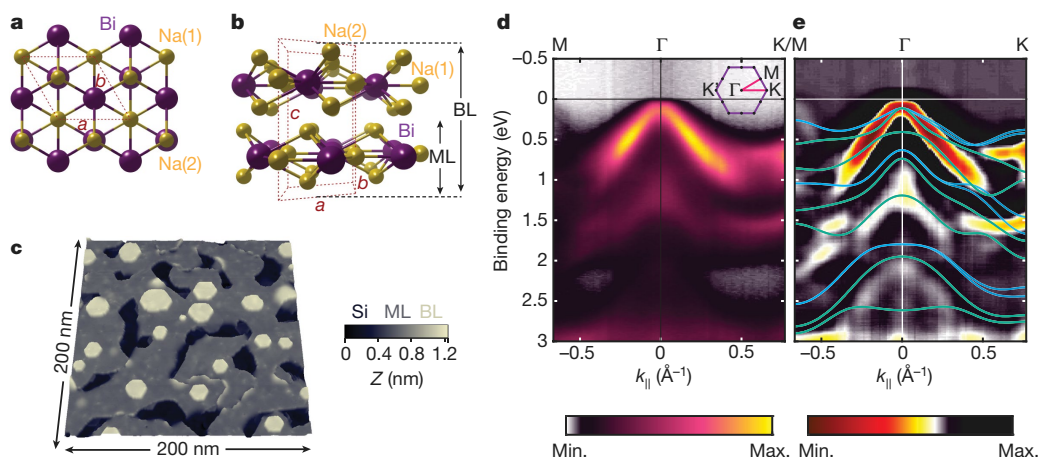


Fig. 1 | Characterization of epitaxial few-layer Na_3Bi and overall electronic structure from ARPES. **a**, **b**, Top (**a**) and side (**b**) view of the 'bulk' Na_3Bi unit cell with $P6_3/mmc$ symmetry, where $a = b = 0.545$ nm and $c = 0.965$ nm. **c**, Large area (200 nm \times 200 nm) topographic STM image (bias voltage $V = 2.0$ V and tunnel current $I = 120$ pA) of few-layer Na_3Bi on Si(111). Regions are colour-coded with ML (monolayer;

dark grey), BL (bilayer; light grey) and Si (Si(111); dark blue). **d**, Overall band structure along the M- Γ -K directions measured with ARPES at $h\nu = 48$ eV, with the intensity of emitted photoelectrons, I , reflected in the colour scale. Inset, 2D Brillouin zone. **e**, Second derivative of **d** (see colour scale) enhances low intensity features. Blue and green curves represent DFT band structures for ML and BL Na_3Bi , respectively.

Fig. 2) demonstrates that the film is electronically 2D, with no dispersion in k_z unlike its bulk or 15-nm thin-film counterparts^{10,15}. Depth-dependent X-ray photoelectron spectroscopy (XPS; see Extended Data Fig. 1) revealed no additional components observed in either the Si 2p core level corresponding to the Si(111) substrate or the Na and Bi core levels of Na_3Bi , verifying that Na_3Bi is free-standing on Si(111).

We first measure the size of the bulk bandgap for ML and BL regions of Na_3Bi by probing the electronic structure with STS, in which the dI/dV spectrum (the differential conductance dI/dV as a function of sample bias V) is proportional to the LDOS at energy $E_F + eV$. Figure 2a shows typical dI/dV spectra for ML (red) and BL (black) with bandgaps corresponding to 0.36 ± 0.025 eV and 0.30 ± 0.025 eV respectively (see Methods for details on extracting bandgap values and discussion on the minimal tip-induced band bending). All dI/dV spectra in Fig. 2a were taken more than 5 nm away from step edges. Figure 2b plots the experimental bandgap (blue squares) in comparison to DFT-calculated values using the generalized gradient approximation (GGA) for pristine Na_3Bi (black circles) and Na_3Bi layers that contain an Na(2) surface vacancy (red circles) (see Methods for details, with associated band structures found in Extended Data Fig. 4). The large bandgap in ML Na_3Bi is consistent with previous calculations¹⁷ and the relatively small change in bandgap from ML to BL observed experimentally is well explained by the DFT calculations that include Na(2) surface vacancies; this vacancy gives rise to a delocalized resonance feature and enhancement of the electronic bandgap¹⁹, resulting in only a small layer-dependent evolution in bandgap. DFT calculations using the GGA method are well known to underestimate the bandgap²⁰, so we employ the more accurate hybrid functional approach with the modified Becke-Johnson (mBJ) potential to better determine the bandgaps for ML Na_3Bi (without vacancies) (see Methods section 'Experiment parameters' for details). This yields a bandgap of 0.43 eV for ML Na_3Bi , compared to the 0.36 eV obtained for GGA. While the GGA value is in excellent agreement with the experimental value of 0.36 ± 0.025 eV, this is probably a coincidence due to a reduction in the experimental bandgap as a result of the electric field effect modulation discussed later. The zero field value is likely to be larger and closer to the mBJ value.

To verify the prediction that ML and BL Na_3Bi are large-bandgap QSH insulators (see Extended Data Fig. 4) we probe the step edge of these islands to the underlying Si(111) substrate to look for the conductive edge state signature of a QSH insulator. STM topography (Fig. 2c) shows a BL Na_3Bi region decorated with Na surface vacancies and an approximately 1.2 nm step edge to the underlying Si substrate, with a small ML Na_3Bi protrusion about 0.7 nm above the substrate. Figure 2d

shows dI/dV spectra for BL Na_3Bi taken 3 nm away from the edge (black curve) and at the edge (blue curve). In contrast to the gap in the bulk, the dI/dV spectrum at the edge is quite different, with states filling the bulk gap along with a characteristic dip at 0 mV bias. Similar features observed in other QSH insulators $1T'-\text{WTe}_2$ (ref. 7) and bismuthene⁶ have been attributed to one-dimensional (1D) non-trivial edge states and the emergence of a Luttinger liquid²¹. Figure 2e shows dI/dV spectra as a function of distance away from the edge, tracing the orange profile in Fig. 2c, demonstrating the extended nature of the edge state feature, with Fig. 2f showing that the average dI/dV signal within the bulk bandgap moving away from the edge follows the expected exponential decay for a 1D topologically non-trivial state⁶.

With ML and BL Na_3Bi verified as large-bandgap QSH insulators, we now examine the role of an electric field in modifying the size and nature of the bandgap. First, we utilize ARPES to measure the band structure after doping the surface with K to generate an electric field. Details on calculating the displacement field are in Methods section 'Mapping from K deposition to electric displacement field'. Figure 3a-d shows the band structure along Γ -K for values of the electric field of 0.0, 0.72, 1.44 and 2.18 V nm^{-1} respectively, with the blue and green/pink dots reflecting the extracted maxima from energy distribution curves (EDC) and momentum distribution curves (MDC; see Methods section 'Extracting and fitting the ARPES band dispersion of few-layer Na_3Bi ' for details). The right panel in each pair in Fig. 3a-d represents a model of a 2D gapped Dirac system (see Methods for details). In Fig. 3a only the hole band is observable, with a hyperbolic band dispersion and asymptotic hole Fermi velocity of $v_F \approx 3 \times 10^5 \text{ m s}^{-1}$. The band dispersion near Γ displays the clear cusp of a band edge indicating a gapped system, with 140 meV separation between the valence band edge and the Fermi energy E_F . The effect of K dosing in Fig. 3b-d is to n-type dope the sample and consequently increase the displacement field. At a displacement field of 0.7 V nm^{-1} the separation from the valence band edge to E_F has increased to about 257 meV. The bandgap must be at least this amount, consistent with STS, though we cannot determine its exact magnitude since the conduction band lies above E_F (although it can be estimated, see Extended Data Fig. 7a). Upon increasing the displacement field, a Dirac-like electron band emerges with asymptotic Fermi velocity $v_F \approx 10^6 \text{ m s}^{-1}$. The weakness in intensity of the conduction band is most likely due to the different orbital characters of the conduction and valence bands (see Extended Data Figs. 4 and 7 for orbitally resolved DFT band structures). At 1.4 V nm^{-1} our best estimate of the gap between the two band edges is about 100 meV and reduces to

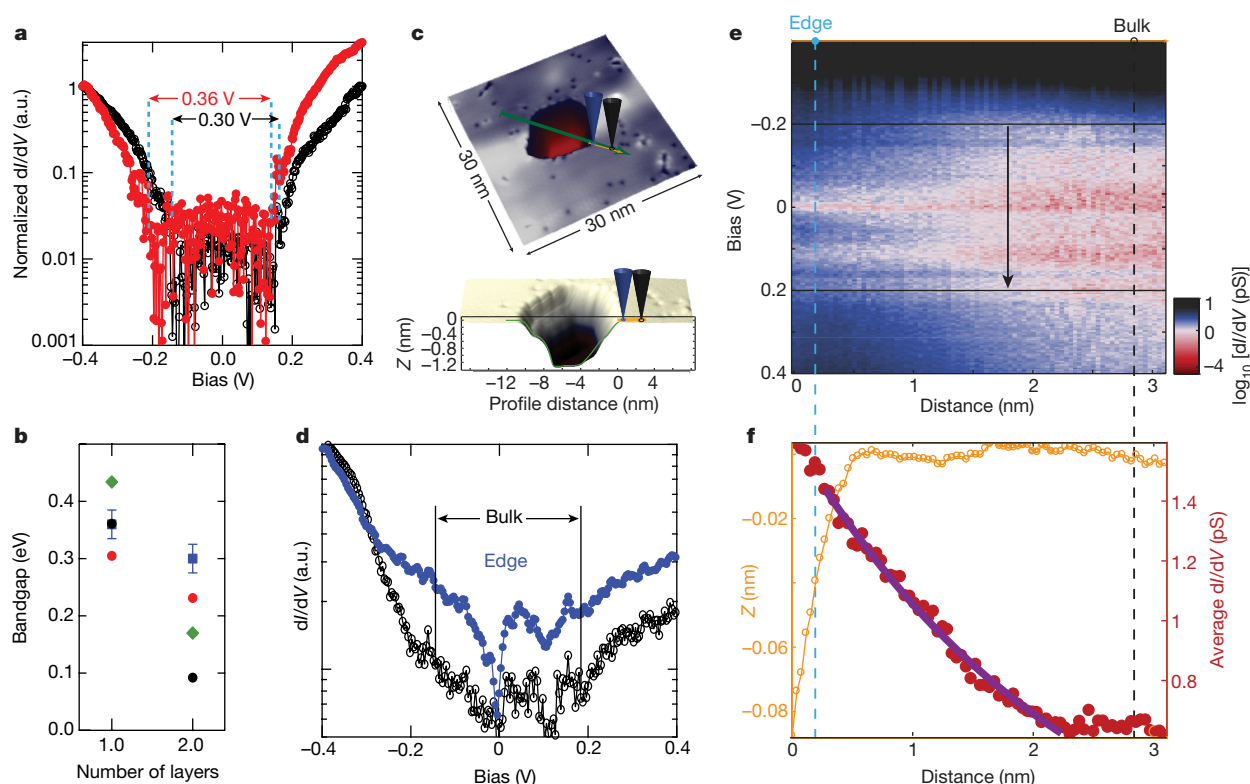


Fig. 2 | Bandgap in ML and BL Na₃Bi and edge state behaviour.

a, Normalized dI/dV spectra corresponding to ML (red) and BL (black) Na₃Bi. a.u., arbitrary units. Conduction and valence band edges are reflected by the sharp onset of dI/dV intensity, marked by dashed blue lines, separated by 0.36 V (ML) and 0.30 V (BL). **b**, Evolution of the bandgap as a function of layer thickness ('number of layers') determined from dI/dV spectra (blue squares, with error bars of ± 25 meV; see Methods for error calculation), DFT calculations using the GGA functional on pristine Na₃Bi (black circles) and with an Na(2) vacancy (red circles), and DFT calculations using the mBJ potential (green diamonds). **c**, Top, STM topography of a region of BL Na₃Bi (grey), and the underlying Si(111) substrate (red). The orange line reflects the region over which the dI/dV measurements were performed in **e**, and the green line extends the

profile across the pinhole step. Bottom, cross-section through the pinhole overlaid with the orange and green line profiles. **d**, dI/dV spectra taken near the step edge of BL Na₃Bi to Si(111) substrate (blue) and in the bulk of BL Na₃Bi (black). **e**, dI/dV colour map (scale at bottom right) taken at and then moving away from the step edge where the dashed vertical lines reflect the spectra shown in **d** and the horizontal lines represent the region over which the dI/dV signal was averaged as shown in **f**. **f**, Corresponding intensity profile of dI/dV in the bulk gap showing the exponential decay away from the step edge, where the orange trace represents the topographic height, Z , and the red points are the average dI/dV magnitude within the bulk bandgap (horizontal dashed region of **e**) fitted to an exponential (thick purple line).

about 90 meV at 2.2 V nm^{-1} (see Methods section 'Electric displacement field dependence from ARPES' for calculation). While a significant reduction in bandgap with displacement field clearly occurs, due to the finite energy width of the bands (approximately 100 meV)

we cannot say definitively whether the gap is fully closed or even reopened again.

To elucidate the effect of an electric field on the electronic structure more clearly, we turn back to measurements made with STM. Here, the

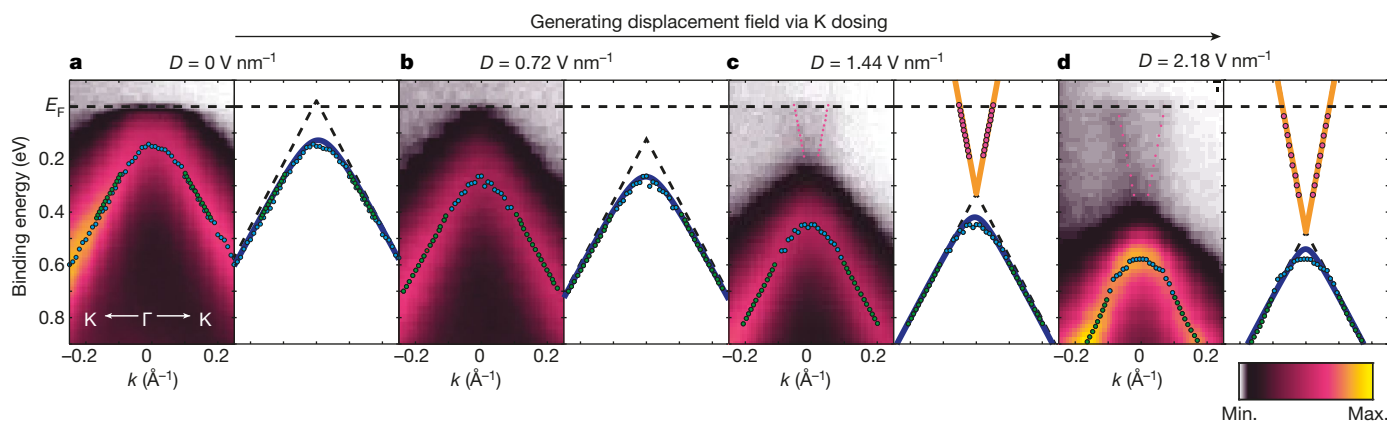


Fig. 3 | Band-structure modulation in ARPES in ML and BL Na₃Bi using potassium (K) dosing. **a–d**, ARPES intensity plots of the band dispersion with K dosing. For each K dose, the left panel shows the ARPES spectra, points are maxima extracted from MDCs (green/pink) and EDCs (blue); right panel shows fits of the maxima to hyperbolas (orange and blue solid lines represent fits of the conduction and valence bands, respectively)

along with linear asymptotes (black dashed lines). **a**, Before K dosing, the hole band is located about 140 meV below E_F ; **b**, after the first K dose, at 0.72 V nm^{-1} displacement field, the hole band is 257 meV below E_F ; **c**, after the second dose, at 1.44 V nm^{-1} displacement field, the electron band is separated from the hole band by about 100 meV; and **d**, after the last dose, at 2.18 V nm^{-1} displacement field the band separation is 90 meV.

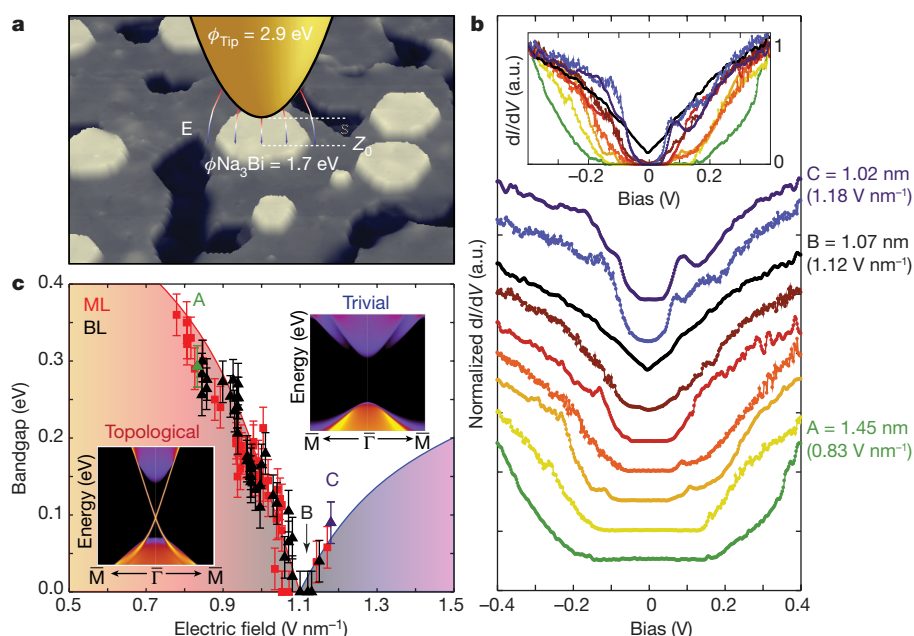


Fig. 4 | Electric-field-induced bandgap modulation of ML and BL Na_3Bi in STM. **a**, Schematic of a metallic tip at distance s above the surface (z_0) of Na_3Bi , with the difference in work function ($\phi_{\text{Tip}} - \phi_{\text{Na}_3\text{Bi}}$) generating a localized electric field (E). **b**, Individual dI/dV spectra taken on BL Na_3Bi at different tip-sample separations (electric fields) as labelled on the figure, where A, B and C correspond to tip heights (electric fields) of 1.45 nm (0.83 V nm^{-1}), 1.07 nm (1.12 V nm^{-1}) and 1.02 nm (1.18 V nm^{-1}), respectively. Spectra have been normalized and offset for

clarity. Inset shows the spectra without an offset. **c**, Bandgap extracted from dI/dV spectra as a function of electric field for ML (red squares) and BL (black triangles) Na_3Bi , with error bars of $\pm 25 \text{ meV}$; see Methods for error calculation. Marked field strengths A, B and C correspond to those indicated in **b**. Orange and purple shaded regions represent guides to the eye. Insets represent DFT projected edge state band structures below and above the critical field, where the colour represents the spectra weight.

tip-sample separation is now varied in order to tune the electric field due to the electrostatic potential difference between the metallic tip and Na_3Bi , as illustrated schematically in Fig. 4a. The electrostatic potential difference is dominated by the difference in work function between the tip and the sample, approximately 1.2 eV (see Methods section ‘Calculating tip-sample separation and electric field’ for calculation). Changes in the bandgap can then be measured in the dI/dV spectra as a function of tip-sample separation and converted to electric field as shown in Fig. 4b (details in Methods). Figure 4b shows normalized dI/dV spectra taken on BL Na_3Bi that are offset for clarity (Extended Data Fig. 9 shows similar spectra taken on ML Na_3Bi) at various tip-sample separations (electric fields). A large modulation occurs upon increasing the electric field strength, with the bandgap reducing from 300 meV to completely closed (and exhibiting the characteristic V-shape of a Dirac semimetal) at about 1.1 V nm^{-1} and then reopening above this to yield a bandgap of about 90 meV at approximately 1.2 V nm^{-1} . Figure 4c plots the bandgap as a function of electric field for ML and BL Na_3Bi , with both exhibiting a similar critical field where the bandgap is closed and then reopened into the trivial/conventional regime with increasing electric field. DFT calculations also predict such a transition, arising from a Stark-effect-induced rearrangement of s - and p -like bands near the gap (see Methods section ‘DFT calculations of electric-field-induced topological phase transition’ and Extended Data Fig. 10 for full calculation). The projected edge state band structures above and below the critical field are shown as insets to Fig. 4c. Note that due to the difficulty of estimating tip-sample distance, the electric field magnitude may include a systematic error as large as 50%, however the trend of gap size with electric field is correct (see Methods for detailed discussion).

By combining ARPES and STS, we have demonstrated that ML and BL Na_3Bi are QSH insulators with bulk bandgaps above 300 meV, offering the potential to support dissipationless transport of charge at

room temperature. An electric field tunes the phase from topological insulator to conventional insulator with a bandgap of about 90 meV due to a Stark-effect-driven transition. This bandgap modulation of more than 400 meV is larger than has been achieved in atomically thin semiconductors such as bilayer graphene^{22,23}, is similar to that achieved in phosphorene²⁴, and may be useful in optoelectronic applications²⁵ in the mid-infrared. Na_3Bi is chemically inert in contact with silicon, and the electric fields required to induce the topological phase transition are below the breakdown fields of conventional dielectrics, meaning that future experiments that measure the ballistic edge current turn on/turn off along the film edges may be possible. These properties make ultrathin Na_3Bi a promising platform for realizing new forms of electronic switches based on topological transistors for low-energy logic circuits.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0788-5>.

Received: 1 March 2018; Accepted: 7 October 2018;

Published online 10 December 2018.

- Qian, X., Liu, J., Fu, L. & Li, J. Quantum spin Hall effect in two-dimensional transition metal dichalcogenides. *Science* **346**, 1344–1347 (2014).
- Liu, J. et al. Spin-filtered edge states with an electrically tunable gap in a two-dimensional topological crystalline insulator. *Nat. Mater.* **13**, 178–183 (2014).
- Pan, H., Wu, M., Liu, Y. & Yang, S. A. Electric control of topological phase transitions in Dirac semimetal thin films. *Sci. Rep.* **5**, 14639 (2015).
- Vandenbergh, W. G. & Fischetti, M. V. Imperfect two-dimensional topological insulator field-effect transistors. *Nat. Commun.* **8**, 14184 (2017).
- König, M. et al. Quantum spin Hall insulator state in HgTe quantum wells. *Science* **318**, 766–770 (2007).
- Reis, F. et al. Bismuthene on a SiC substrate: a candidate for a high-temperature quantum spin Hall material. *Science* **357**, 287–290 (2017).
- Tang, S. et al. Quantum spin Hall state in monolayer $1T'\text{-WTe}_2$. *Nat. Phys.* **13**, 683–687 (2017).
- Fei, Z. et al. Edge conduction in monolayer WTe_2 . *Nat. Phys.* **13**, 677–682 (2017).

9. Wu, S. et al. Observation of the quantum spin Hall effect up to 100 kelvin in a monolayer crystal. *Science* **359**, 76–79 (2018).
10. Liu, Z. K. et al. Discovery of a three-dimensional topological Dirac semimetal, Na_3Bi . *Science* **343**, 864–867 (2014).
11. Wang, Z. et al. Dirac semimetal and topological phase transitions in A_3Bi ($\text{A}=\text{Na}, \text{K}, \text{Rb}$). *Phys. Rev. B* **85**, 195320 (2012).
12. Liu, Z. K. et al. A stable three-dimensional topological Dirac semimetal Cd_3As_2 . *Nat. Mater.* **13**, 677–681 (2014).
13. Borisenko, S. et al. Experimental realization of a three-dimensional Dirac semimetal. *Phys. Rev. Lett.* **113**, 027603 (2014).
14. Hellerstedt, J. et al. Electronic properties of high-quality epitaxial topological Dirac semimetal thin films. *Nano Lett.* **16**, 3210–3214 (2016).
15. Zhang, Y. et al. Molecular beam epitaxial growth of a three-dimensional topological Dirac semimetal Na_3Bi . *Appl. Phys. Lett.* **105**, 031901 (2014).
16. Schumann, T. et al. Observation of the quantum Hall effect in confined films of the three-dimensional Dirac semimetal Cd_3As_2 . *Phys. Rev. Lett.* **120**, 016801 (2018).
17. Niu, C. et al. Robust dual topological character with spin-valley polarization in a monolayer of the Dirac semimetal Na_3Bi . *Phys. Rev. B* **95**, 075404 (2017).
18. Wu, J. et al. High electron mobility and quantum oscillations in non-encapsulated ultrathin semiconducting $\text{Bi}_2\text{O}_2\text{Se}$. *Nat. Nanotechnol.* **12**, 530–534 (2017).
19. Edmonds, M. T. et al. Spatial charge inhomogeneity and defect states in topological Dirac semimetal thin films of Na_3Bi . *Sci. Adv.* **3**, eaao6661 (2017).
20. Perdew, J. P. et al. Understanding band gaps of solids in generalized Kohn-Sham theory. *Proc. Natl Acad. Sci. USA* **114**, 2801–2806 (2017).
21. Voit, J. One-dimensional Fermi liquids. *Rep. Prog. Phys.* **58**, 977–1116 (1995).
22. Ohta, T., Bostwick, A., Seyller, T., Horn, K. & Rotenberg, E. Controlling the electronic structure of bilayer graphene. *Science* **313**, 951–954 (2006).
23. Zhang, Y. et al. Direct observation of a widely tunable bandgap in bilayer graphene. *Nature* **459**, 820–823 (2009).
24. Deng, B. et al. Efficient electrical control of thin-film black phosphorus bandgap. *Nat. Commun.* **8**, 14474 (2017).
25. Ju, L. et al. Tunable excitons in bilayer graphene. *Science* **358**, 907–910 (2017).

Acknowledgements M.T.E. was supported by ARC DECRA fellowship DE160101157. M.T.E., J.L.C., C.L. and M.S.F. acknowledge funding support

from CE170100039. J.L.C., J.H. and M.S.F. are supported by M.S.F.'s ARC Laureate Fellowship (FL120100038). S.A. acknowledges funding support from ARC Discovery Project DP150103837. M.T.E. and A.T. acknowledge travel funding provided by the International Synchrotron Access Program (ISAP) managed by the Australian Synchrotron, part of ANSTO, and funded by the Australian Government. M.T.E. acknowledges funding from the Monash Centre for Atomically Thin Materials Research and Equipment Scheme. S.A.Y. and W.W. acknowledge funding from Singapore MOE AcRF Tier 2 (grant no. MOE2015-T2-2-144). S.A. acknowledges the National University of Singapore Young Investigator Award (R-607-000-094-133). This research used resources of the Advanced Light Source, which is a DOE Office of Science User Facility under contract no. DE-AC02-05CH11231. Part of this research was undertaken on the soft X-ray beamline at the Australian Synchrotron, part of ANSTO. The authors acknowledge computational support from the National Supercomputing Centre, Singapore.

Author contributions M.T.E., J.L.C. and M.S.F. devised the STM experiments. M.T.E. devised the ARPES and XPS experiments. M.T.E. and J.L.C. performed the MBE growth and STM/STS measurements at Monash University. J.H. assisted with the experimental setup at Monash University. M.T.E., J.L.C. and A.T. performed the MBE growth and ARPES measurements at Advanced Light Source with the support from H.R., S.T. and S.-K.M. The MBE growth and XPS measurements at the Australian Synchrotron were performed by M.T.E., J.L.C., A.T., J.H. and C.L. The DFT calculations were performed by L.C.G., J.N.B.R., W.W. and S.A.Y.; S.A. assisted with the theoretical interpretation of the data. M.T.E., J.L.C. and M.S.F. composed the manuscript. All authors read and contributed feedback to the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0788-5>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.T.E.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Growth of few-layer Na₃Bi on Si(111). Ultrathin Na₃Bi thin films were grown in ultrahigh vacuum (UHV) molecular beam epitaxy (MBE) chambers and then immediately transferred after the growth under UHV to the interconnected measurement chamber (that is, Createc LT-STM at Monash University, Scientia R-4000 analyser at Advanced Light Source and SPEC Phoibos 150 at Australian Synchrotron). To prepare an atomically flat substrate, a p-type Si(111) wafer was flash annealed to 1,250 °C in order to achieve a (7 × 7) surface reconstruction (this was achieved via direct current heating at Monash University and the Australian Synchrotron, and via electron bombardment heating at the Advanced Light Source).

For Na₃Bi film growth, effusion cells were used to simultaneously evaporate elemental Bi (99.999%) in an overflux of Na (99.95%) with a Bi:Na flux ratio not less than 1:10, calibrated by quartz microbalance. During growth, the substrate temperature was held between 300 °C and 330 °C to ensure successful crystallization. At the end of the growth, the sample was left at the growth temperature for 10 min in a Na overflux to improve film quality before cooling to room temperature.

Experiment parameters. STM/STS measurements. These were performed in a Createc LT-STM operating at 5 K. A PtIr tip was prepared and calibrated using an Au(111) single crystal, confirming the presence of the Shockley surface state at −0.5 V and flat LDOS near the Fermi level before all measurements. After measurements on few-layer Na₃Bi were completed, the Au(111) sample was re-measured to confirm the tip had not significantly changed and still exhibited flat LDOS near the Fermi level. STM differential conductance measurements (dI/dV) were measured using a 5 mV AC excitation voltage (673 Hz) that was added to the tunnelling bias. Differential conductance measurements were made under open feedback conditions with the tip in a fixed position above the surface. Data were prepared and analysed using MATLAB and IGOR Pro.

ARPES measurements. These were performed at Beamline 10.0.1 of the Advanced Light Source (ALS) at Lawrence Berkeley National Laboratory, USA. Data were taken using a Scientia R4000 analyser at 20 K sample temperature. The total energy resolution was 20–25 meV depending on the beamline slit widths and analyser settings, and the angular resolution was 0.2°. This resulted in an overall momentum resolution of ~0.01 Å^{−1} for photoelectron kinetic energies measured, with the majority of the measurements performed at $h\nu = 48$ eV and 55 eV. Doping of the Na₃Bi films was performed via in situ K dosing from a SAES getter source in UHV. Dosing was performed at a temperature of 20 K to avoid K clustering.

XPS measurements. These were performed at the Soft X-ray Beamline of the Australian Synchrotron using a SPECS Phoibos-150 spectrometer at room temperature. The Bi 5d and Na 2p of the Na₃Bi, as well as the Si 2p of the Si(111) substrate, were measured at photon energies of 100 eV, 350 eV, 800 eV and 1,487 eV. This ensured surface sensitivity for the low photon energy scans at 100 eV, with the higher photon energies used to characterize the depth dependence of the core levels, in particular whether there was any chemical bonding between the Na₃Bi film and the Si(111) substrate. The binding energy scale of all spectra are referenced to the Fermi energy (E_F), determined using either the Fermi edge or 4f core level of an Au reference foil in electrical contact with the sample. Core level spectra were analysed using a Shirley background subtraction. Kelvin probe measurements to determine the work function of Na₃Bi were also performed on this system.

Density-functional theory calculations. First-principles calculations based on density-functional theory (DFT) were used to obtain electronic band structures of monolayer (ML) and bilayer (BL) Na₃Bi, with and without Na(2) vacancies. This was achieved using the projector augmented wave (PAW) method²⁶ with calculations implemented in Quantum ESPRESSO code and the Vienna *ab initio* Simulation Package (VASP). The generalized gradient approximation (GGA) using the Perdew–Burke–Ernzerhof (PBE) functional²⁷ for the exchange–correlation potential was adopted. The plane-wave cutoff energy was set to be 400 eV. The Brillouin zone sampling was performed by using k grids with a spacing of $2\pi \times 0.02$ Å^{−1} within a Γ -centred sampling scheme. For all calculations, the energy and force convergence criteria were set to be 10^{-5} eV and 10^{-2} eV Å^{−1}, respectively. For the ML and BL Na₃Bi calculations with Na surface vacancies there is 1 Na vacancy per 2 × 2 supercell. For the Na₃Bi layers, we used a vacuum region of thickness greater than 15 Å to eliminate the artificial interaction between the periodic images. The edge states were studied by constructing the maximally localized Wannier functions²⁸ and by using the iterative Green's function method²⁹ as implemented in the WannierTools package³⁰. More accurate calculations of the bandgap for ML and BL Na₃Bi (without Na vacancies) were performed via a more accurate hybrid functional approach using the modified Becke–Johnson potential^{31–33}.

Characterization via RHEED, LEED and XPS. In order to confirm that the few-layer films of Na₃Bi are high quality and epitaxial over large areas, we performed reflection high-energy electron diffraction (RHEED), low energy electron diffraction (LEED) and X-ray photoelectron spectroscopy (XPS), as shown in Extended Data Fig. 1. Extended Data Fig. 1a shows the characteristic RHEED pattern for Si(111) 7 × 7 reconstruction along $\Gamma - \bar{M}$, while Extended Data Fig. 1b shows the

RHEED pattern for few-layer Na₃Bi along $\Gamma - \bar{K}$, consistent with RHEED reported on films of 15 unit cell thickness, where the lattice orientation of Na₃Bi is rotated 30° with respect to the Si(111) substrate¹⁵. Extended Data Fig. 1c shows the 1 × 1 LEED pattern consistent with growth of Na₃Bi in the (001) direction. The sharpness of the spots and the absence of rotational domains indicates high-quality single crystal few-layer Na₃Bi over a large area. Extended Data Fig. 1d shows the Bi 5d and Na 2p core levels of a few-layer Na₃Bi film taken at $h\nu = 100$ eV, with the peak positions consistent with published results on 20 nm film and bulk Na₃Bi^{10,15,34}.

To rule out reaction of Na₃Bi with the Si substrate, we performed depth-dependent XPS (by varying the photon energy in order to increase the kinetic energy of emitted photoelectrons, as a result increasing the mean free path) to examine the Na₃Bi–Si(111) interface. The Na and Bi core levels exhibited no additional components (data not shown). Extended Data Fig. 1e shows XPS of the Si 2p core level (reflecting the substrate) at photon energy $h\nu = 350$ eV (left panel) and $h\nu = 850$ eV (right panel). In each panel the black curve represents the Si 2p core level of the bare Si(111) 7 × 7 and the red curve represents the Si 2p core level with few-layer Na₃Bi grown on top. In each case, the spectra have been normalized to the maximum in intensity and energy-corrected (to account for the small interfacial charge transfer that occurs) in order to overlay the core levels. The spectra have been offset for clarity. It is clear there is negligible change to the Si 2p core level after Na₃Bi growth, with no additional components or significant broadening arising, verifying that Na₃Bi is free-standing on Si(111). This is consistent with the fact that our ARPES measurements on ultrathin Na₃Bi showed no features with the Si(111) 7 × 7 symmetry.

Band dispersion in k_z from photon-energy-dependent ARPES. Photon energy-dependent ARPES can be used to determine whether a material possesses a 3D band dispersion, that is, the binding energy E_B depends not only on in-plane wavevectors k_x and k_y , but also on out-of-plane wavevector k_z . To determine the momentum perpendicular to the surface requires measuring energy distribution curves as a function of the photon energy in order to measure E_B versus k_z , using the nearly free-electron final state approximation^{35,36}.

$$k_z = \sqrt{\frac{2m}{\hbar^2} (E_k + V_0 - E_k \sin^2 \theta)} \quad (1)$$

where θ is the emission angle, m is the effective mass of an electron, V_0 is the inner potential (reflecting the energy difference between the bottom of the valence band and the vacuum level) and E_k is the kinetic energy of the emitted photoelectrons (where $E_k = h\nu - \Phi - E_B$, with $h\nu$ the photon energy, Φ the work function and E_B the energy). At normal emission (that is, $\theta = 0$), equation (1) simplifies to:

$$k_z = \sqrt{\frac{2m}{\hbar^2} (E_k + V_0)} \quad (2)$$

Therefore, using equation (2) and measuring energy distribution curves at normal emission as a function of photon energy, we can directly measure E_B versus k_z assuming an inner potential $V_0 = 12.5$ eV for Na₃Bi determined in ref. 10.

Extended Data Fig. 2 shows a colour plot of k_z as a function of binding energy (and reflects energy distribution curves taken at normal emission for photon energies between 45 eV and 55 eV. A flat band is observed near 0 eV (the Fermi energy) and represents the valence band maximum. This band possesses no dispersion in k_z (that is, no bulk band dispersion), verifying that few-layer Na₃Bi is indeed electronically 2D, unlike its thin-film and bulk counterparts.

Bandgap extraction from STM spectra. Determining the bulk electronic bandgap of ML and BL Na₃Bi was achieved by performing scanning tunnelling spectroscopy (dI/dV as a function of sample bias V) more than 5 nm away from step edges. The valence and conduction band edges in the LDOS are defined as the onset of differential conductance intensity above the noise floor. Owing to the large variation in dI/dV signal near a band edge, it is difficult to determine the bandgap on a linear scale as shown in Extended Data Fig. 3a, and it is therefore useful to plot the logarithm of the dI/dV curves for accurate bandgap determination, as shown in Extended Data Fig. 3b.

For measurements involving tuning the electric field by varying the tip–sample distance, dI/dV spectra were taken over a wide range of tunnelling currents (0.01–1 nA), resulting in large changes in signal at band edges and a change in the relative magnitude of signal to noise. In order to unambiguously determine the magnitude of the gap without reference to the noise magnitude, we adopted the following procedure. Spectra were normalized to a relatively featureless point in the LDOS away from the band edge onset. The dI/dV signal corresponding to a bias of −400 meV was chosen for normalization. After the normalization procedure was completed for all spectra, we take the band edges as the point at which the dI/dV has fallen to 0.01 of the normalized value. We find that this definition closely corresponds to the onset of conductance above the noise floor (Extended Data Fig. 3c). Normalization was also performed at −300 meV, 300 meV and 400 meV with only a small variation (<15 meV) observed, due to the sharp onset

in conductance in both the valence and conduction bands. Accounting for error in the normalization and the tip-induced band bending discussed below yields an error in determining the gap magnitude of ± 25 meV.

Tip-induced band bending. Tip-induced band bending (TIBB) effects have the potential to overestimate the size of the electronic bandgap due to unscreened electric fields and can strongly influence the interpretation of STM data. The absence/presence of TIBB is usually verified by performing dI/dV spectra at different initial current setpoints (different tip-sample separations). In the absence of TIBB there will be negligible change in the band edges of the spectra; however, if the spectra are strongly influenced by TIBB, increasing the current setpoint (reducing tip-sample separation) will lead to increased band bending, and overestimation of the band-gap³⁷. As shown in Fig. 4b the exact opposite is observed for few-layer Na_3Bi . In this case the bandgap becomes smaller, closes and then re-opens upon increasing the current (and consequently electric field), and is clearly not consistent with TIBB. However, we also cannot rule out whether the dI/dV spectra taken at low currents (low fields) (as shown in Fig. 2) are intrinsic or free from TIBB.

In order to estimate the effects of TIBB, we first adopt a model based on a uniformly charged sphere³⁷, with the analytic expression for the difference ϕ_{BB} between the apparent (measured) and actual energy position of a spectral feature (for example, band edge) due to TIBB being given by:

$$\phi_{\text{BB}}(V_b, r, h, \varepsilon) = \frac{1}{1 + \varepsilon \frac{h}{r}} (eV_b - W_0) \quad (3)$$

where ε is the dielectric constant, V_b is the bias voltage, h the tip height, r the tip radius and W_0 the work function difference between sample and tip (that is, $W_0 = W_{\text{sample}} - W_{\text{tip}}$). We use $h = 1.5$ nm and $W_0 = -1.2$ eV (taken from calculations in Methods section ‘Calculating tip-sample separation and electric field’) and assume a tip radius $r = 25$ nm. For the static dielectric constant, we use the value¹⁴ for bulk Na_3Bi , $\varepsilon = 120$. While the dielectric constant in ML and BL Na_3Bi could be very different, the bulk value is currently the only available value and is used as a rough estimate. Extended Data Fig. 3d plots ϕ_{BB} as a function of bias voltage calculated from equation (3). The correction factor for TIBB is given by $1 / \left(1 - \frac{\partial \phi}{\partial V}\right)$, which yields 1.13 for Na_3Bi .

The uniformly charged sphere model is known to overestimate TIBB by almost a factor of 2 when more detailed modelling that incorporates charge redistribution is taken into account³⁷. Therefore, we adopt the image charges method for a charged sphere in front of a dielectric sample (see Appendix from ref.³⁷). In this case the TIBB expression is replaced by:

$$\phi_{\text{BB}}(V_b, r, h, \varepsilon) = F(r, h, \varepsilon) (eV_b - W_0) \quad (4)$$

where F is the ratio of the electrical potential on the tip surface and at the point of the sample closest to the tip. For $h = 1.5$ nm, $r = 25$ nm and $\varepsilon = 120$ we find that $F = 0.064$, yielding a TIBB correction factor of 1.07, meaning our best estimate is that the measured bandgap of ultrathin Na_3Bi includes a systematic overestimation due to TIBB of 7%. For ML and BL Na_3Bi with bandgaps of 360 meV and 300 meV this corresponds to 25 meV and 21 meV respectively. Because this is comparable to the random error of ± 25 meV as discussed in Methods section ‘Bandgap extraction from STM spectra’, we have not corrected for TIBB.

DFT calculations for Na(2) vacancy Na_3Bi layers and Z2 calculations. The calculated band structures of Na_3Bi ML and BL with an Na(2) vacancy including spin-orbit coupling (SOC) are displayed in Extended Data Fig. 4b and g, respectively. Bandgaps of 0.30 eV (0.28 eV) and 0.22 eV (0.16 eV) are obtained for ML (without structural relaxation) and BL respectively, with the bandgaps obtained from DFT using the Quantum Espresso code and the value obtained from the VASP package in brackets. The non-trivial topological character can be intuitively observed from the orbital components of the band edge states. Without SOC (as shown for ML and BL in Extended Data Fig. 4c and h respectively), the conduction band minimum (CBM) is mainly contributed by Na s and Bi s orbitals, whereas the valence band maximum (VBM) is mainly from the Bi p_x/p_y orbitals, showing a normal band ordering. After including SOC (shown in Extended Data Fig. 4d and i for ML and BL respectively), the band ordering is inverted at Γ , with p orbitals at the CBM above the s orbitals at the VBM. This SOC-induced band inversion marks a topologically non-trivial phase, which indicates that both ML and BL Na_3Bi are nontrivial 2D topological insulators. The edge state spectrum is shown in Extended Data Fig. 4e for ML and in Extended Data Fig. 4j for BL. The projected 1D Brillouin zone is shown in Extended Data Fig. 4a. One can clearly observe a Kramers pair of topological edge states. To prove these systems are non-trivial, we determine the topological invariant of both systems. This is done by employing the Wilson loop method^{38,39}, in which one traces the evolution of the Wannier function centres, as plotted in Extended Data Fig. 1f and k. From the calculation, we confirm that both ML and the BL Na_3Bi with and without Na(2) vacancies are topologically non-trivial with the invariant $\mathbb{Z}_2 = 1$.

Extracting and fitting the ARPES band dispersion of few-layer Na_3Bi . Energy dispersion curves (EDCs) and momentum dispersion curves (MDCs) are slices through constant momentum and constant energy of the photoemission spectra (such as Extended Data Fig. 5a) along high-symmetry directions ($M-\Gamma-M$) or ($K-\Gamma-K$). Band energy and momentum coordinates are extracted by Gaussian fitting of the photoemission intensity on a flat background (as shown in Extended Data Fig. 5b, c by the blue circles). We find that band edges are extracted more reliably from EDCs, while MDC peak positions are used at larger binding energies where clearly distinct peaks can be resolved (see left panel of Extended Data Fig. 5b).

The measured bands are observed to fit hyperbolae as shown in Extended Data Fig. 5d, as expected of gapped Dirac systems. Fits of the bands to a parabola are much poorer, as shown in Extended Data Fig. 5e. In order to accurately model the band dispersion of a 2D gapped Dirac system, we use a bi-partite model for the valence (p) and conduction (n) bands that assumes the form:

$$(E_{B,i} - D)^2 = \Delta_i^2 + \hbar^2 v_{F,i}^2 (k + k_0)^2, i \in p, n \quad (5)$$

where $\Delta = \Delta_n + \Delta_p$ represents the energy gap, $v_{F,i}$ the asymptotic Fermi velocities at large momenta, E_B the binding energy, and D a doping or energy-shift of the bands.

The velocities $v_{F,i}$ measured for the valence and conduction bands for both films are nearly independent of K dosing, with near-isotropic dispersion in k_x, k_y (also shown in Extended Data Fig. 6). Hence we take $v_{F,n}$ and $v_{F,p}$ to be a global fit parameter, with best fit values $v_{F,n} \approx 1 \times 10^6$ m s⁻¹ and $v_{F,p} \approx 3 \times 10^5$ m s⁻¹.

We then fit the valence band photoemission—that is, the negative solution for equation (5)—using the global $v_{F,p}$ parameter, allowing us to determine Δ_p and D as a function of K dosing. A monotonic increase of D with K dosing is observed as expected, reflecting the shift of the valence band to larger binding energy.

The photoemission intensity of the electron band is two orders of magnitude less than that of the valence band—possibly due to the different orbital characters of the two bands resulting in a lower intensity due to matrix element effects. Owing to the large bandgap of few-layer Na_3Bi , the conduction band lies well above the Fermi level in the as-grown film, meaning that significant charge transfer from K dosing is needed to n-type dope the film in order to observe the conduction band. As such the fitting parameter Δ_n for the electron bands can only be determined once the conduction band is resolvable below E_F and in addition further seen to match the valence band determined value for D . Values for $\Delta_{n,p}$ as a function of electric displacement field are addressed in Methods section ‘Electric displacement field dependence from ARPES’.

Mapping from K deposition to electric displacement field. Potassium deposited on the Na_3Bi surface donates electrons leaving a positive K^+ ion behind, producing a uniform planar charge density. This is equivalent to a parallel plate capacitor, allowing the electric displacement field, D , to be calculated across the Na_3Bi film using Gauss’ law via:

$$D = \frac{E}{\varepsilon} = \frac{eQ}{\varepsilon_0} = \frac{e\Delta n}{\varepsilon_0} \quad (6)$$

with Q representing the total charge transferred due to potassium doping, that is, $Q = \Delta n$. The charge transfer to the system cannot be directly inferred when a Fermi surface cannot be clearly resolved, so our calculations make use of the conduction band Fermi surface that becomes distinct after 15 min of K dosing. As seen in Extended Data Fig. 6a, the n-type Fermi surface is a nearly isotropic Dirac cone. By measuring k_F as a function of K dosing either from EDCs or a Fermi surface map as in Extended Data Fig. 6b, the charge density can be directly calculated using:

$$n(k_F) = \frac{g}{4\pi} k_F^2 \quad (7)$$

where a band degeneracy of $g = 4$ can be taken for Dirac systems¹⁴. The charge density $n(k_F)$ is also consistent with the assumption of a Dirac dispersion centred at D , that is, $k_F = D/\hbar v_F$.

The change in $n(k_F)$ as a function of K dosing is approximately 2×10^{12} cm⁻² between consecutive K dosing until the 50-min mark (where charge saturation occurs). By assuming that in this regime every K atom donates one electron and a constant dose rate we can extrapolate the total $n(k_F)$ back to the doping of the as-grown film growth. For the as-grown film this corresponds to a p-type doping of 4×10^{12} cm⁻². From this as-grown doping we can then calculate the electric displacement field using equation (6), as shown on the right hand axis of Extended Data Fig. 6c.

Electric displacement field dependence from ARPES. Next we map the calculated $\Delta_{n,p}$ (which reflects the size of the bandgap) at different K dosing to a corresponding electric displacement field, as shown in Extended Data Fig. 7a. The purple circles in Extended Data Fig. 7a represent where $\Delta_{n,p}$ are directly extracted

from the bi-partite model of the experimental data. At low displacement field, where the conduction band is still above the Fermi level, we cannot directly obtain a value for Δ_n . We estimate the size of Δ_n at these low displacement fields using the ratio $(\Delta_p + \Delta_n)/\Delta_p \approx 1.4$, which is directly obtained from the purple points. For the as-grown sample this yields a value of ~ 320 meV, which, given we cannot directly measure the conduction band edge, is in reasonable agreement with the experimental result from STS, and the theoretical DFT value. The reduction in bandgap is consistent with the independently measured gap-closing from STS in Fig. 4c, with the relative energy separation of the electron and valence bands narrowing monotonically with increasing field. When the estimated bandgap $\Delta = \Delta_n + \Delta_p$ becomes as small as 100 meV (corresponding to displacement fields > 1.5 V nm $^{-1}$), it is comparable to the intrinsic energy broadening of the bands (particularly the valence band). Therefore, we cannot definitely conclude from ARPES measurements whether the bandgap completely closes after this point or is re-opened again.

Calculating tip-sample separation and electric field. In STM for a simple square barrier the tunnelling current (in atomic units) follows:

$$I_t = I_0 e^{-2(z-z_0)\sqrt{2\Phi}} \quad (8)$$

where Φ is the work function of the energy barrier and z and z_0 the tip and sample positions, respectively, such that the distance between tip and sample is $s = z - z_0$. This allows the work function of the barrier to be obtained by measuring the tunnelling current as a function of tip position, then extracting the slope of $\ln(I_t)/z$:

$$\frac{d \ln(I)}{dz} = -2\sqrt{2\Phi} \quad (9)$$

Extended Data Fig. 8 shows logarithmic plots of the tunnelling current as a function of relative distance for Au(111) (bias 500 mV) (Extended Data Fig. 8a) and thin-film Na₃Bi (bias -300 mV) where the absolute current is plotted (Extended Data Fig. 8b). For Au(111) the characteristic exponential dependence of $I(z)$ (straight line on the semi-log plot) is observed with current increasing from 0.01 nA to 10 nA by moving the tip 3 Å closer to the surface. However, it is immediately clear that very different behaviour occurs for Na₃Bi in Extended Data Fig. 8b. At low tunnelling current an exponential dependence with distance $I(z)$ is observed, but as the distance from tip and sample decreases the current saturates to a value around 1 nA, which occurs over a length scale of ~ 1 nm. This corresponds to the barrier height going to zero as the tip approaches the sample surface.

This effect results from a modification of equation (8) due to the lowering of the potential barrier by the mirror potential seen by an electron in close proximity to a metal surface⁴⁰, and is most pronounced for low-work-function materials, such as Na₃Bi. This is shown schematically in Extended Data Fig. 8c. As we show in detail below, the theoretical treatment indicates that the effect of the mirror potential at large distances is simply a rigid shift in distance of the region in which the exponential behaviour occurs. Thus, we can use the exponential region where equation (8) is obeyed to determine the work function, and we can extrapolate equation (8) to point contact to determine the tip-sample distance, provided a correction factor is applied to account for the rigid shift due to the mirror potential.

There has been significant work on modelling image potential effects, well summarized in ref. ⁴¹, whereby considering a simple model such as the square barrier depicted in Extended Data Fig. 8c for long tip-sample distances, $\ln(2)/s \gg \Phi$, the tunnelling current can be expressed as:

$$I_t = I_0 e^{-2z\sqrt{2\Phi}} e^{\ln(2)\sqrt{2\Phi}} \quad (10)$$

where the main effect of the image potential at very long tip-sample distances is to increase the tunnelling current by a constant factor $e^{\ln(2)\sqrt{2\Phi}}$. For $I_t = I_0$, we obtain the increased distance between tip and sample due to the image potential $s = \ln(2)/2\Phi$ in atomic units. Converting from atomic units to SI units gives

$$\begin{aligned} s(\text{\AA}) &= \left(\frac{\ln(2)}{2\Phi(\text{eV})} \right) \times 1 \text{ Bohr}(\text{\AA}) \times 1 \text{ Hartree}(\text{eV}) \\ &= \left(\frac{\ln(2)}{2\Phi(\text{eV})} \right) \times 0.529 \text{\AA} \times 27.2 \text{ eV} \end{aligned}$$

such that equation (8) would underestimate z_0 by a distance of $5 \text{\AA}/[\Phi(\text{eV})]$.

This value, however, is most probably an underestimate, as shown in Extended Data Fig. 8d (taken directly from figure 10 of ref. ⁴¹) where $\ln(I(s)/I_0)$ as a function of distance is plotted for the simple square barrier model described above (dashed line), and more sophisticated calculations incorporating DFT-LDA (full line) and a model developed by Pitarke et al., that accounts for both non-local and local exchange and correlation effects (dashed-dotted red line)⁴². At large s the models still show an exponential $I(s)$ with similar slope (Extended Data Fig. 8d) but the

distance at which similar current is achieved is increased by as much as a factor of 3 from the square barrier model, such that the underestimation of z_0 distance by equation (8) is more likely to be $\sim 15 \text{\AA}/[\Phi(\text{eV})]$ ⁴².

To understand and quantify the tip-sample distance in Na₃Bi, we first extract the slope from the linear region in Extended Data Fig. 8b (as shown by the black line). This fit yields a slope of 1.5\AA^{-1} , and accounting for the bias of -300 mV yields a work function or barrier height of 2.3 eV. I versus z measurements (where z is the nominal distance between tip and sample as defined by the STM piezo controller with arbitrary zero) were taken at various negative and positive bias values and yielded work functions of 2.3 ± 0.05 eV. This confirms a very low barrier height. From the above analysis, we estimate that the effect of the mirror potential is to lead to an underestimation of z_0 in equation (8) of $\sim 15 \text{\AA}/2.3$ or $\sim 6.5 \text{\AA}$.

In order to determine the exact value of z_0 , we assume that at $z_0 = 0$ (point contact) the conductance is of the order of the conductance quantum $e^2/h \approx 40 \mu\text{S}$, which gives $I_0 = 12 \mu\text{A}$ at a bias voltage of -300 mV. Then the fit to the exponential region of $I(z)$ in Extended Data Fig. 8b to equation (10) gives $z_0 = -13.2 \text{\AA}$. This allows us to plot the absolute distance $s = z - z_0$ on the top axis of Extended Data Fig. 8b.

We are now in a position to estimate the tip-sample separation from the dI/dV measurements. For each dI/dV curve we extract a relative distance s for each dI/dV measurement by referencing the tunnelling current at -300 mV to the I versus s data in Extended Data Fig. 8b.

The green and purple dI/dV spectra in Fig. 4b are treated in the following manner:

Green curve. At a bias of -300 mV this has a tunnelling current of 26 pA. From the I versus z plot taken at -300 mV this corresponds to relative z distance of 1.3\AA . Adding 13.2\AA to account for point contact and mirror potential yields a tip-sample separation of 14.5\AA .

Purple curve. At a bias of -300 mV this has a tunnelling current of 570 pA. From the I versus z plot taken at -300 mV this corresponds to a relative z distance of -3\AA . Adding 13.2\AA to account for point contact and mirror potential yields a tip-sample separation of 10.2\AA .

We now have an estimate of the tip-sample separation, so to calculate a displacement field for each of our dI/dV measurements we need to calculate the potential difference, that is, the work function difference between the metallic tip and the few-layer Na₃Bi. The measured barrier height of 2.3 eV is an average of the tip and sample work functions, $\Phi_{\text{Barrier}} = (\Phi_{\text{Tip}} + \Phi_{\text{Na3Bi}})/2$, meaning we need to know either the tip work function or measure the work function of the Na₃Bi.

The Kelvin probe technique was used to measure the work function of few-layer Na₃Bi. The work function was determined by measuring the contact potential difference of the Na₃Bi relative to a gold reference of known work function (determined by photoelectron spectroscopy secondary electron cutoff measurements). A work function for few-layer Na₃Bi of 1.7 ± 0.05 eV was measured using this technique. This value and the 2.3 eV potential barrier gives a tip work function of 2.9 ± 0.05 eV, and a potential difference between tip and sample of 1.2 ± 0.1 eV. It should be noted that this is significantly lower than the expected value for a PtIr tip, suggesting that Na atoms have been picked up by the tip and consequently lower the work function (work function of Na is 2.23 eV). While Na atoms were picked up they have little influence on the tip density of states, as spectroscopy performed on Au(111) after measurements on Na₃Bi were completed revealed a flat LDOS near the Fermi energy.

Using the calculated potential difference and tip-sample separation from above allows us to calculate electric fields for dI/dV spectra on ML and BL Na₃Bi, as shown in Fig. 4b for BL and in Extended Data Fig. 9 for ML, with the full data set of bandgap as a function of electric field shown in Fig. 4c.

The above discussion reflects our best estimate of the electric field magnitude. However, we estimate that the electric field could be incorrect by as much as 50% due to the main source of error in estimating the correction for the mirror potential. An estimate of this can be made by comparing the correction we employed and the correction based on the square barrier model which is 3 times smaller. This $\sim 4 \text{\AA}$ smaller correction would lead to a large increase in electric field, $\sim 50\%$. In addition, there are small sources of error that arise from the error in the Na₃Bi work function and in the fit of the slope of $\ln(I)/dz$, the potential difference and barrier height. However, these errors are expected to be no more than 10%. Regardless, the qualitative behaviour—that is, closing and re-opening of the bandgap with increasing electric field—is correct and independent of knowing the exact electric field magnitude.

DFT calculations of electric-field-induced topological phase transition. In order to confirm our predicted experimental observation of a topological phase transition, DFT calculations were performed on ML Na₃Bi with an Na(2) surface vacancy as a function of electric field, as shown in Extended Data Fig. 10. Extended Data Fig. 10a plots the bandgap variation as a function of electric field for ML Na₃Bi with an Na(2) vacancy. The critical field where the bandgap closes and

then reopens is $1.85 \text{ V } \text{\AA}^{-1}$. This value is an order of magnitude larger than the experimental result of $\sim 1.1 \text{ V nm}^{-1}$. A possible explanation is that in the modelling there is 1 Na vacancy per 2×2 supercell (an order of magnitude greater than the defect density observed experimentally). These charged defects would be expected to induce stronger screening of external electric fields, which tends to make the critical electric field overestimated compared to experiment.

Importantly, the DFT results clearly demonstrate the topological phase transition, as shown in Extended Data Fig. 10c–f. In Extended Data Fig. 10c, at $0 \text{ V } \text{\AA}^{-1}$, the orbital resolved band structure clearly shows a band inversion of the s and p atomic orbitals at Γ induced by SOC. This clearly indicates a non-trivial 2D topological insulator, as shown in the corresponding projected edge spectrum in Extended Data Fig. 10d with the observation of topological edge states. In Extended Data Fig. 10e, the orbital resolved band structure above the critical field value where the gap has reopened has undergone a band ordering change at Γ as compared with Extended Data Fig. 10c. This indicates a topological to trivial insulator phase transition with electric field, and is confirmed by the disappearance of the topological edge states in Extended Data Fig. 10f.

Summary of crystalline symmetries for ML and BL Na_3Bi . The preservation and breaking of discrete symmetries under the effects of an out-of-plane E -field, as well as the presence of an $\text{Na}(2)$ vacancy, are tabulated in Extended Data Fig. 11a and b for ML and BL Na_3Bi , respectively.

Notes on nomenclature. (1) C_{nx} and C_{nz} are rotations of $360^\circ/n$ with the rotation axis along the x direction and the z direction, respectively. (2) S_{nz} is a rotation of $360^\circ/n$ followed by reflection in a plane perpendicular to the rotation axis (along the z direction). (3) M_{xy} , M_{xz} and M_{yz} are mirror reflections with the mirror in the x – y plane, the x – z plane and the y – z plane, respectively. (4) I is inversion symmetry. (5) Symmetry relations: $S_{3z} = C_{3z}M_{xy}$; $M_{xz} = C_{2x}M_{xy}$; $S_{6z} = IC_{3z}$; $M_{yz} = IC_{2x}$.

Na_3Bi monolayer (ML). For ML without an $\text{Na}(2)$ vacancy, the space group is $P\bar{6}m2$ (no. 187). The point group is D_{3h} . Both electric field and $\text{Na}(2)$ vacancy break the horizontal mirror symmetry, details are summarized in Extended Data Fig. 11a. Extended Data Fig. 11 shows the Na_3Bi crystal structure for pristine ML (Extended Data Fig. 11c), with an Na surface vacancy (Extended Data Fig. 11d) and with an Na surface vacancy and electric field (Extended Data Fig. 11e).

Na_3Bi bilayer (BL). For BL without an $\text{Na}(2)$ vacancy, the space group is $P\bar{3}m1$ (no. 164). The point group is D_{3d} . Both electric field and $\text{Na}(2)$ vacancy break the inversion symmetry, details are summarized in Extended Data Fig. 11b. Extended Data Fig. 11 shows the Na_3Bi crystal structure for pristine BL (Extended Data Fig. 11f), BL with an Na surface vacancy (Extended Data Fig. 11g) and BL with an Na surface vacancy and electric field (Extended Data Fig. 11h).

One can observe that both BL and ML Na_3Bi possess C_{3z} and C_{2x} rotational symmetries. The difference is that ML has mirror symmetry M_{xy} and BL has inversion symmetry I . In terms of symmetry breaking, the electric field does not break any additional symmetry that is not already broken by the $\text{Na}(2)$ vacancy. For ML, both electric field and $\text{Na}(2)$ vacancy break the horizontal mirror symmetry and C_{2x} rotation symmetry, and reduce the D_{3h} symmetry to C_{3v} symmetry; while for BL, they both break the inversion symmetry and C_{2x} rotation symmetry, and reduce

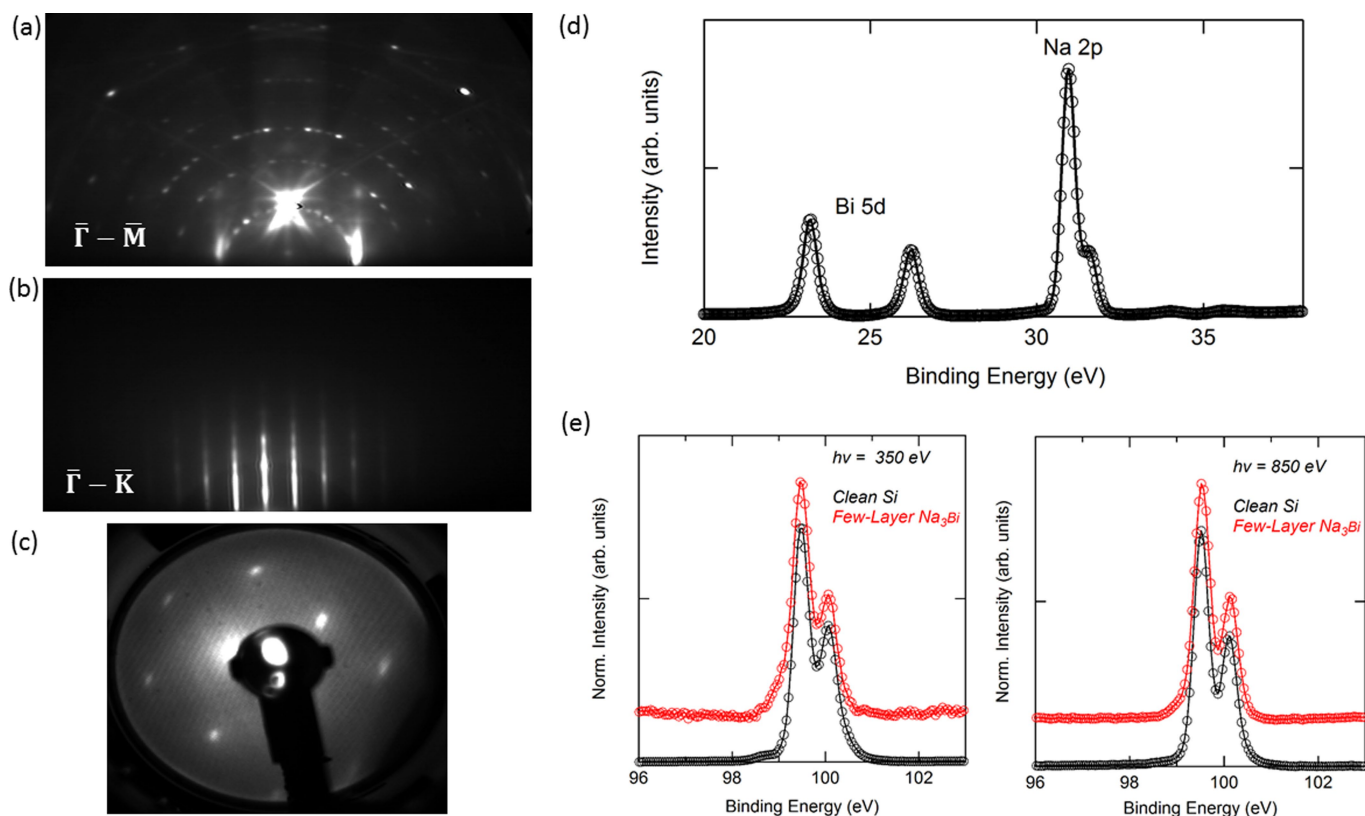
the D_{3d} symmetry to C_{3v} symmetry. Hence, ML and BL share the same point group C_{3v} , after symmetry breaking by electric field or $\text{Na}(2)$ vacancy.

However, for our discussion on the band topology, the crystalline symmetry is not relevant. This is because our main point is that in the absence of electric field, the ML/BL Na_3Bi (with or without Na vacancies) are QSH insulators, which only require time reversal symmetry. Electric field does not break the time reversal symmetry, rather it drives a QSH insulator $v \in \mathbb{Z}_2 = 1$ to trivial insulator $v \in \mathbb{Z}_2 = 0$ transition through a band ordering inversion due to the Stark effect.

Data availability

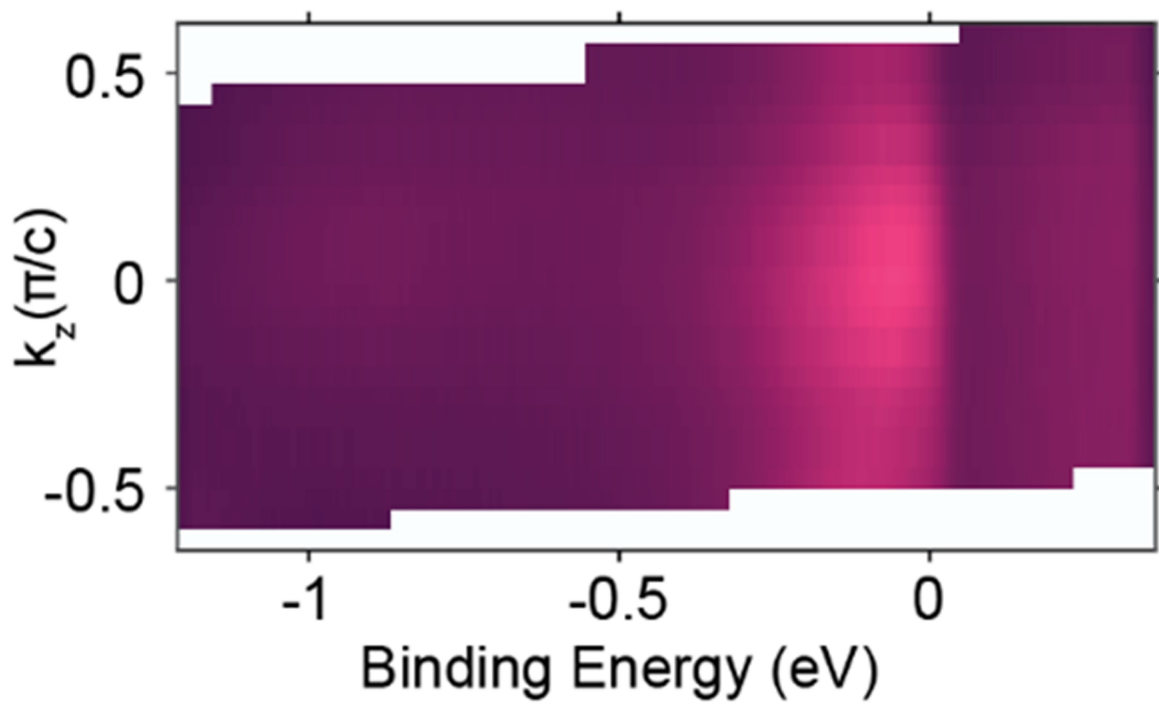
The data that support the findings of this study are available from the corresponding author upon reasonable request.

26. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
27. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
28. Souza, I., Marzari, N. & Vanderbilt, D. Maximally localized Wannier functions for entangled energy bands. *Phys. Rev. B* **65**, 035109 (2001).
29. Lopez Sancho, M. P., Lopez Sancho, J. M., Sancho, J. M. L. & Rubio, J. Highly convergent schemes for the calculation of bulk and surface Green functions. *J. Phys. F* **15**, 851–858 (1985).
30. Wu, Q., Zhang, S., Song, H. F., Troyer, M. & Soluyanov, A. A. WannierTools: An open-source software package for novel topological materials. *Comput. Phys. Commun.* **224**, 405–416 (2018).
31. Becke, A. D. & Johnson, E. R. A simple effective potential for exchange. *J. Chem. Phys.* **124**, 221101 (2006).
32. Tran, F. & Blaha, P. Accurate band gaps of semiconductors and insulators with a semilocal exchange–correlation potential. *Phys. Rev. Lett.* **102**, 226401 (2009).
33. Koller, D., Tran, F. & Blaha, P. Improving the modified Becke–Johnson exchange potential. *Phys. Rev. B* **85**, 155109 (2012).
34. Edmonds, M. T., Hellerstedt, J., O'Donnell, K. M., Tadich, A. & Fuhrer, M. S. Molecular doping the topological Dirac semimetal Na_3Bi across the charge neutrality point with F4-TCNQ. *ACS Appl. Mater. Interfaces* **8**, 16412–16418 (2016).
35. Damascelli, A., Hussain, Z. & Shen, Z.-X. Angle-resolved photoemission studies of the cuprate superconductors. *Rev. Mod. Phys.* **75**, 473–541 (2003).
36. Hufner, S. *Photoelectron Spectroscopy: Principles and Applications* 3rd edn (Springer, Berlin, 2003).
37. Battisti, I. et al. Poor electronic screening in lightly doped Mott insulators observed with scanning tunneling microscopy. *Phys. Rev. B* **95**, 235141 (2017).
38. Yu, R., Qi, X. L., Bernevig, A., Fang, Z. & Dai, X. Equivalent expression of Z2 topological invariant for band insulators using the non-abelian Berry connection. *Phys. Rev. B* **84**, 075119 (2011).
39. Soluyanov, A. A. & Vanderbilt, D. Wannier representation of Z2 topological insulators. *Phys. Rev. B* **83**, 035108 (2011).
40. Lang, N. D. Apparent barrier height in scanning tunnelling microscopy. *Phys. Rev. B* **37**, 10395–10398 (1988).
41. Blanco, J. M., Flores, F. & Perez, R. STM-theory: image potential, chemistry and surface relaxation. *Prog. Surf. Sci.* **81**, 403–443 (2006).
42. Pitarke, J. M., Echenique, P. M. & Flores, F. Apparent barrier height for tunneling electrons in STM. *Surf. Sci.* **217**, 267–275 (1989).



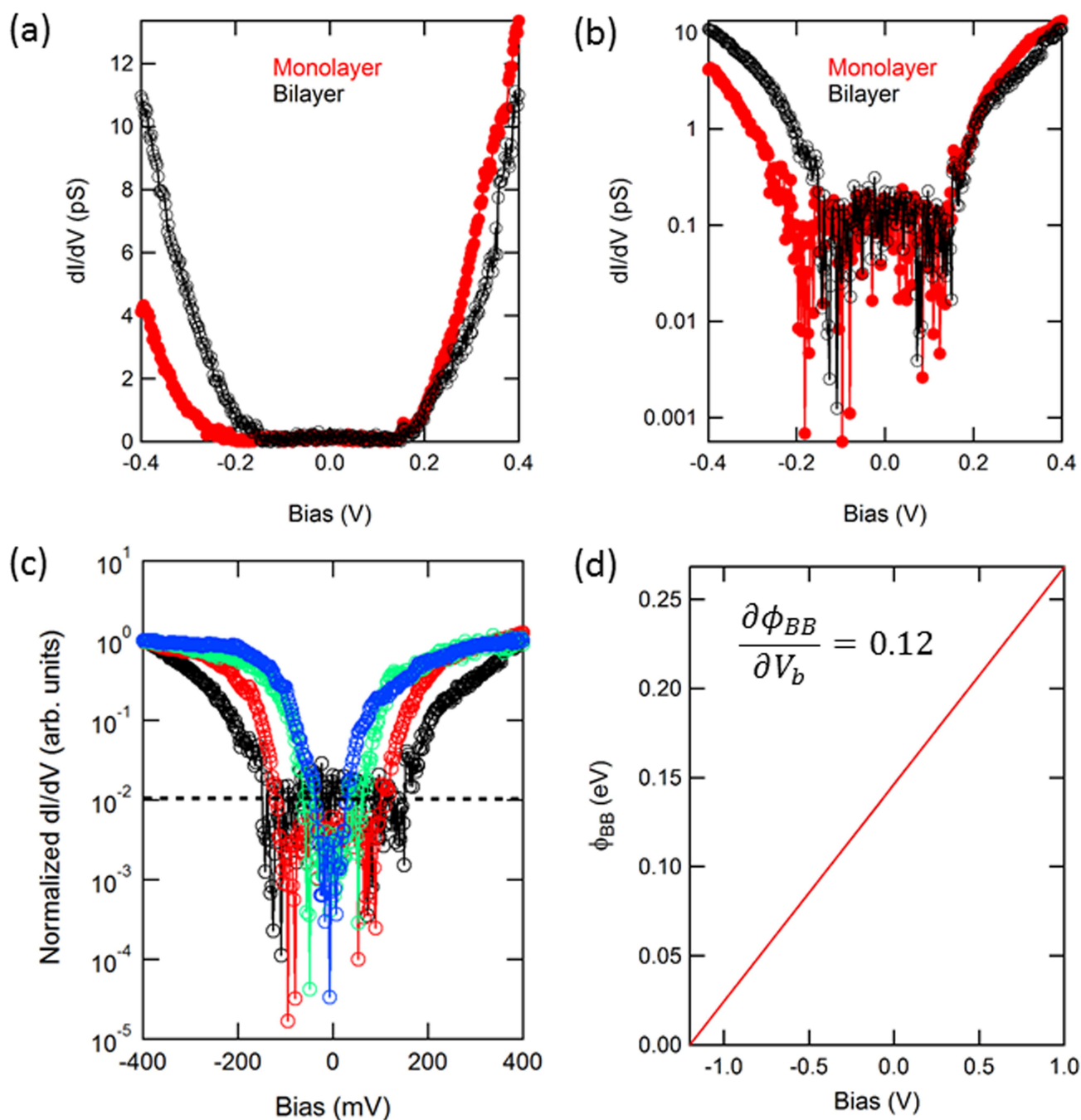
Extended Data Fig. 1 | RHEED, LEED and XPS characterization of few-layer Na_3Bi . **a, b**, RHEED patterns of Si(111) 7×7 reconstruction along the $\bar{\Gamma}-\bar{M}$ direction (**a**) and of few-layer Na_3Bi along the $\bar{\Gamma}-\bar{K}$ direction (**b**). **c**, 1×1 LEED image of few-layer Na_3Bi taken at 32 eV. **d**, XPS of Na 2p and Bi 5d core level taken at $h\nu = 100$ eV for few-layer

Na_3Bi . **e**, Normalized XPS of Si 2p core level taken at 350 eV (left panel) and at 850 eV (right panel). Each panel shows the Si 2p of the clean Si substrate (black curve) and with few-layer Na_3Bi grown on top (red curve). The spectra have been offset in intensity for clarity.



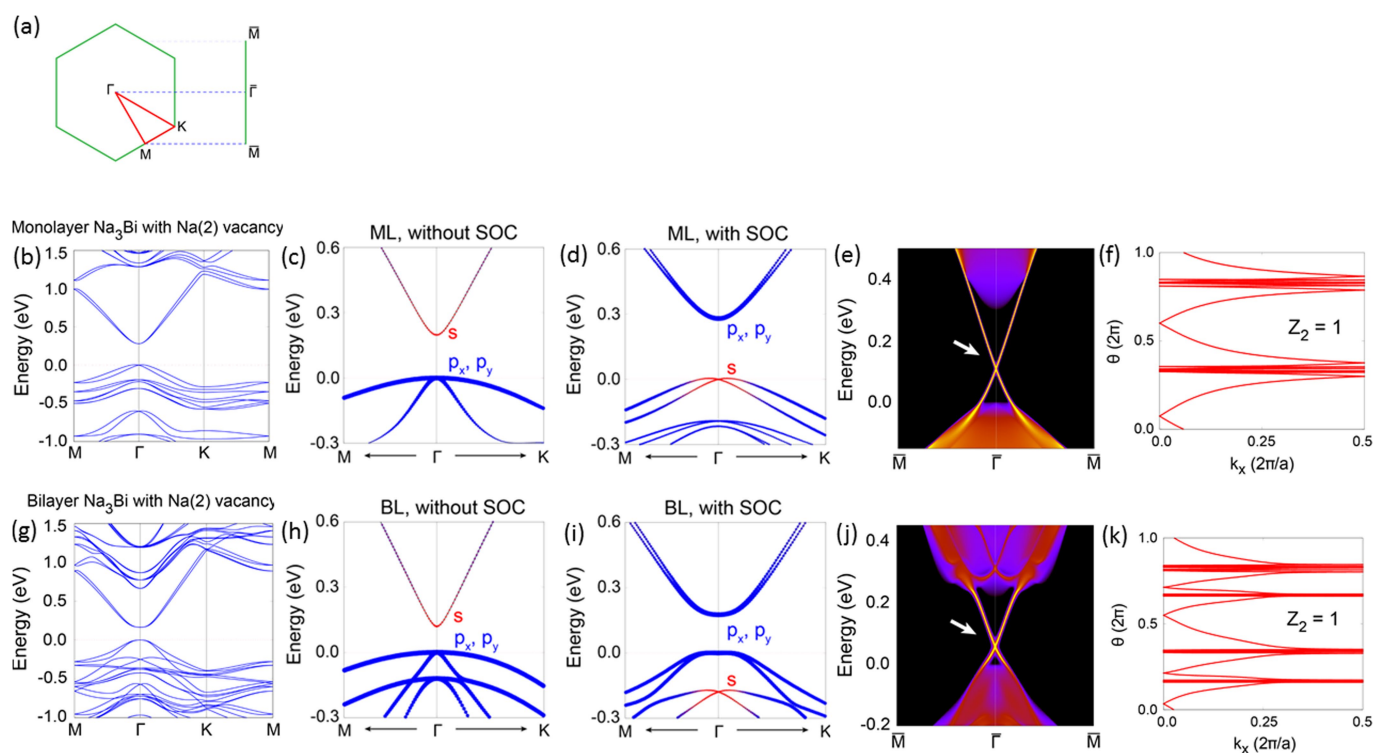
Extended Data Fig. 2 | k_z band dispersion of few-layer Na_3Bi from photon-energy-dependent ARPES measurements. These photon-energy-dependent ARPES measurements of ML/BL Na_3Bi demonstrate effectively 2D dispersion. The figure shows a cut through the photon-

energy-dependent ($h\nu = 45\text{--}55$ eV) Fermi surface, demonstrating non-dispersion of the gapped Dirac valence band along k_z , with the intensity of emitted photoelectrons reflected in the colour scale.



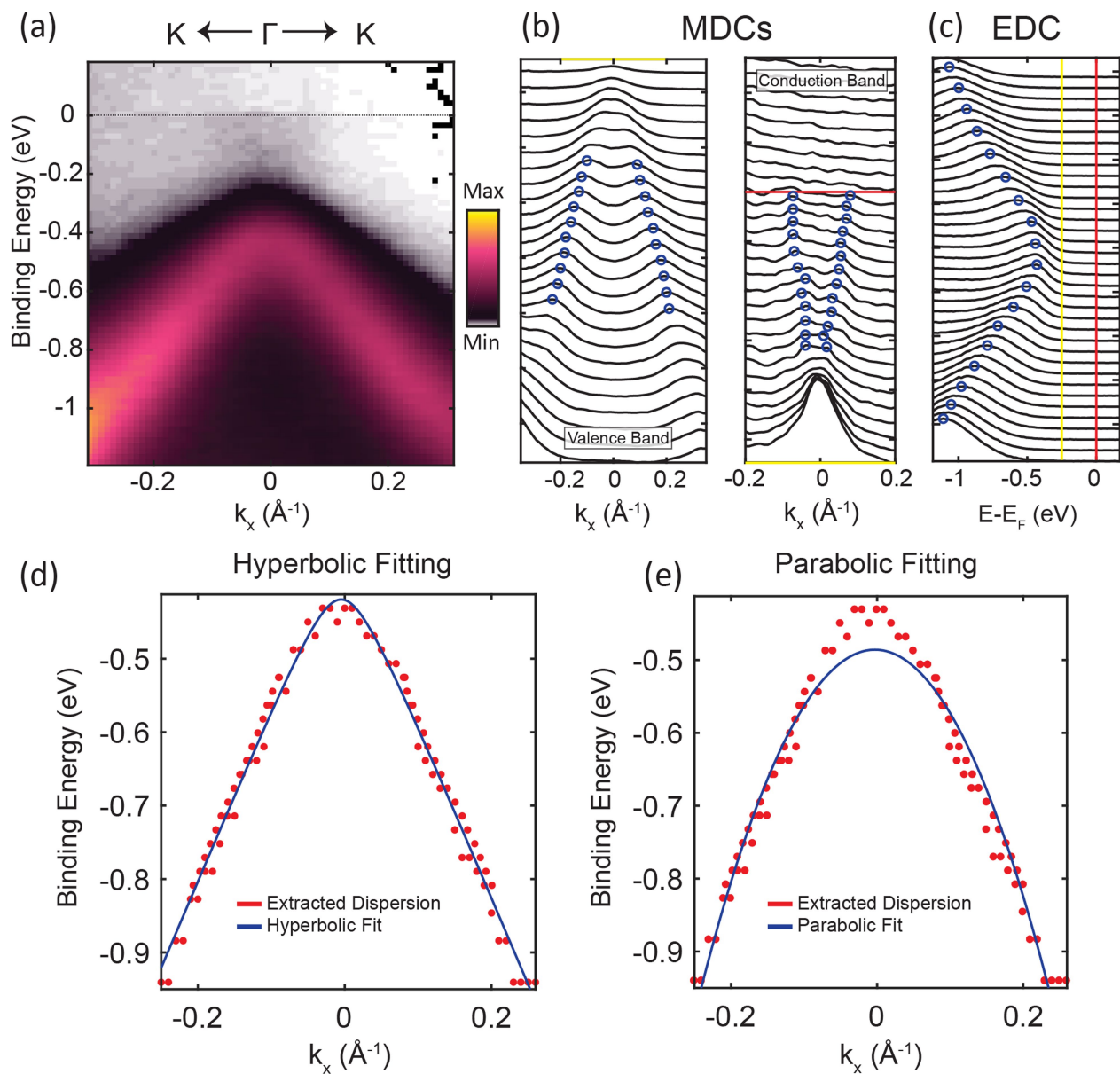
Extended Data Fig. 3 | Bandgap extraction from scanning tunnelling spectra. **a, b**, dI/dV spectra taken for ML (red) and BL (black) Na_3Bi plotted on a linear (**a**) and a logarithmic (**b**) scale. The logarithmic scale better accounts for the large change in intensity near the band edge. **c**, Normalized dI/dV spectra at various tip-sample separations, illustrating

that the onset in intensity typically occurs at a normalized dI/dV signal of 0.01 within an error of ± 25 meV (when also accounting for tip-induced band bending). **d**, Calculated tip-induced band bending, ϕ_{BB} , using equation (3) for different biases.



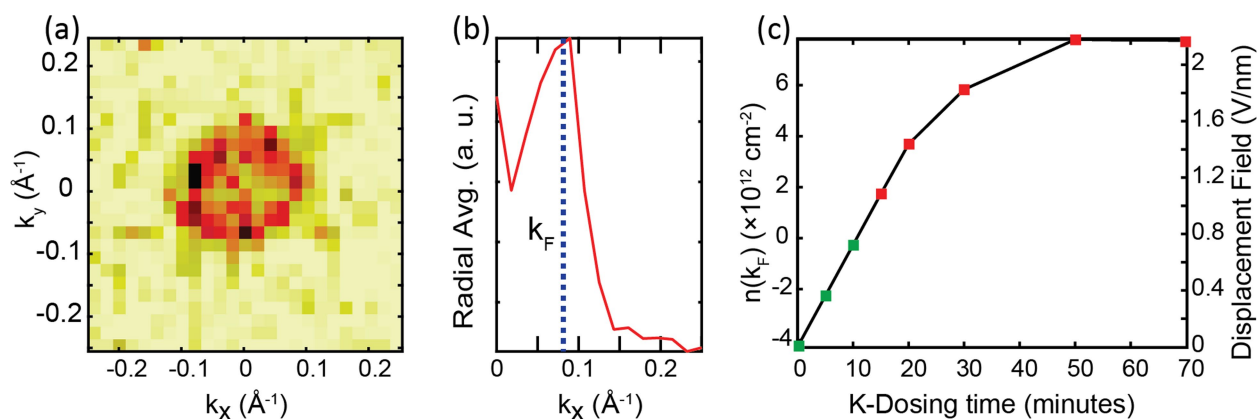
Extended Data Fig. 4 | DFT calculations on the topological nature of Na_3Bi layers. **a**, 2D Brillouin zone for Na_3Bi layered structures. Here we also show the projected 1D Brillouin zone used for studying the edge spectrum. **b–h**, Results for ML (**b–f**) and BL (**g–k**) Na_3Bi with Na(2) vacancies (with one Na(2) vacancy in a 2×2 supercell). **b**, **g**, Electronic band structures, where the energy zero is set to be at the valence band maximum. **c**, **h**, Orbitally resolved band structure without SOC and **d**, **i**, orbitally resolved band structure with SOC. The red dots represent the contribution from the Na s and Bi s atomic orbitals, and the blue dots

represent contribution from the Bi p_x/p_y atomic orbitals. Band inversion induced by SOC can be clearly observed at the Γ point for both ML and BL cases, which indicates that both ML and BL Na_3Bi are non-trivial 2D topological insulators. **e**, **j**, Projected edge spectrum (edge along the $[010]$ direction), where pairs of \mathbb{Z}_2 topological edge states can be observed in the energy gaps (marked by the white arrows) where the colour scale reflects the spectral weight. **f**, **k**, Calculated Wannier function centre evolutions, which indicate a nontrivial \mathbb{Z}_2 invariant ($\mathbb{Z}_2 = 1$) for the bulk band structure in both ML and BL.



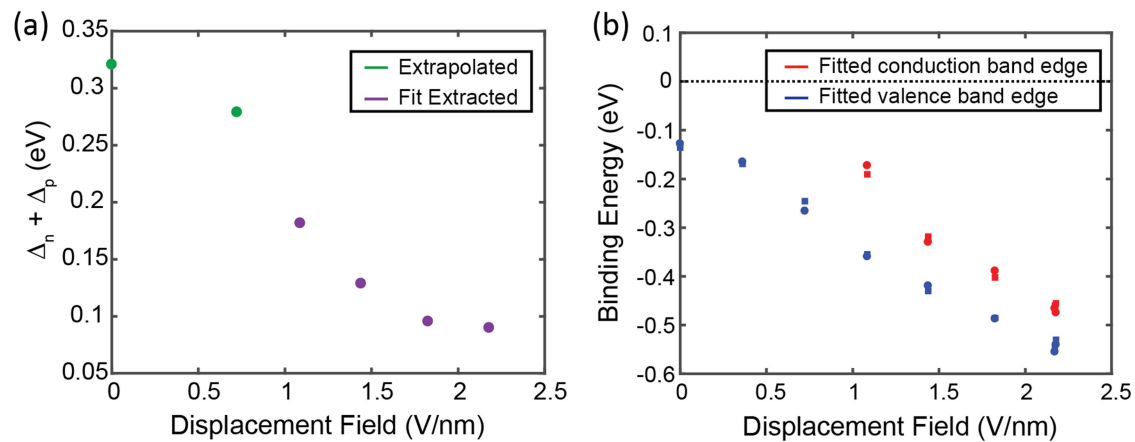
Extended Data Fig. 5 | Extracting and fitting the dispersion relation from ARPES MDC and EDC spectra of few-layer Na_3Bi . **a**, ARPES intensity plot (colour scale at right) along the $\text{K}-\Gamma-\text{K}$ direction after 30 min of K dosing. **b**, Stack plots of MDCs for the valence band (left panel) and conduction band (right panel) extracted from **a**, where the blue

circles represent the Gaussian-fit band locations. Red and yellow lines mark shared constant-energy k_x vertices used throughout **b** and **c**. **c**, EDCs extracted from **a**. **d**, **e**, Fitting band coordinates that have been extracted by MDC and EDC analysis (red dots) to a hyperbola (**d**) and to a parabola (**e**), showing that Na_3Bi -like bands are best described by a hyperbola function.



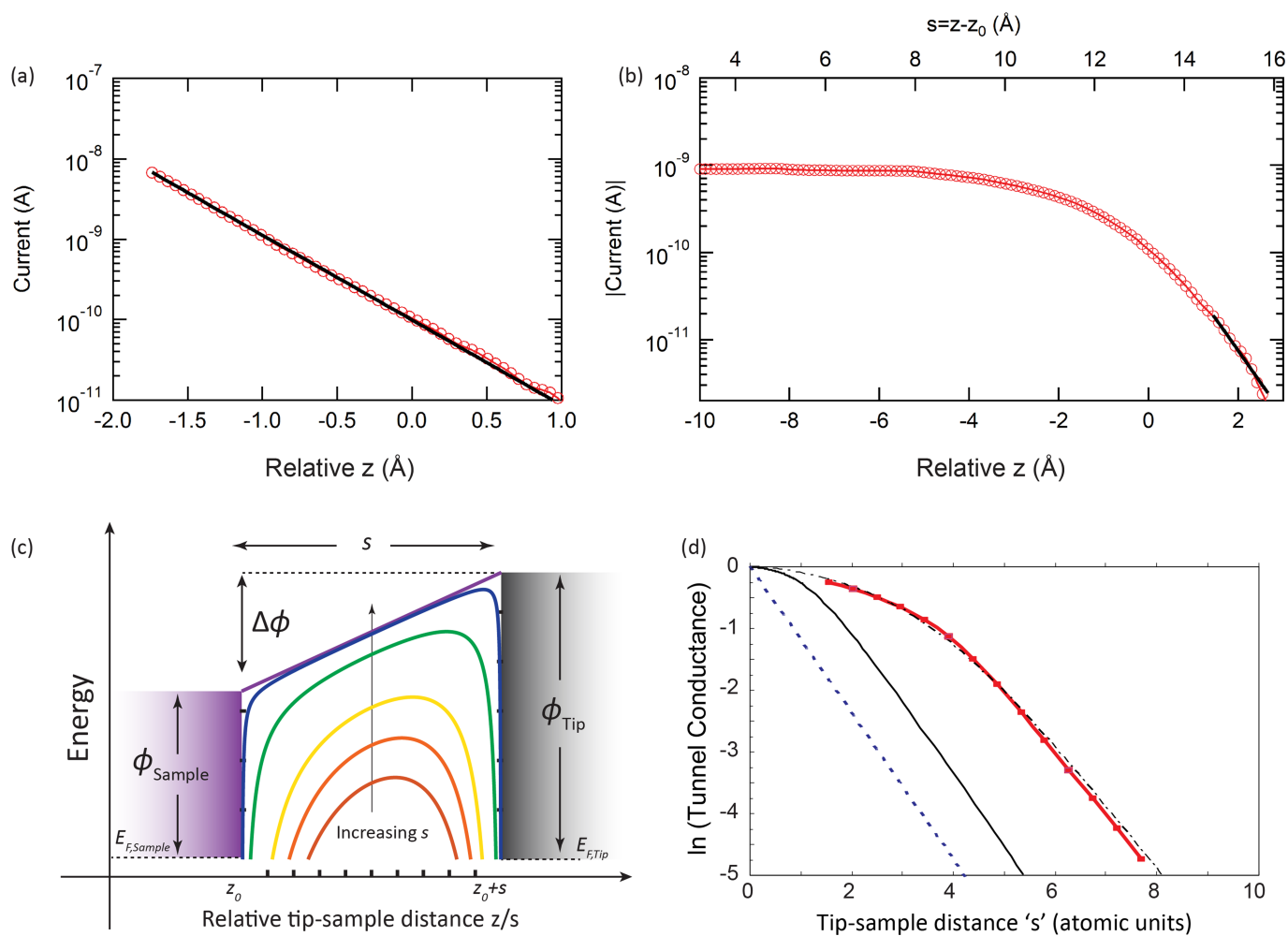
Extended Data Fig. 6 | Displacement field calculations from K dosing in ARPES. **a**, Electron-band Fermi surface of few-layer Na_3Bi after 30 min of K dosing, where the colour scale reflects the intensity of emitted photoelectrons at the Fermi energy. **b**, Red line, radially averaged momentum profile through the Fermi surface, showing the ring structure at k_F (dashed blue line shows position of k_F). **c**, Calculated charge density

(left-hand y axis) using equation (7) and the corresponding electric displacement field (right-hand y axis) associated with the net charge transfer from the as-grown film as a function of K-dosing time. Red points are as-measured, and green points are extrapolated based on the E_F shifting rate with K dosing between 15 min and 50 min.



Extended Data Fig. 7 | Electric displacement field-dependence of topological insulator ML/BL Na₃Bi bands near E_F . **a**, The sum of Δ parameters from the best fit of equation (5) to ARPES dispersion versus applied electric displacement field. Both $\Delta_{n,p}$ are directly calculated from the high-field (purple) measurements, however in low-field (green) measurements, E_F is not sufficiently shifted for the electron dispersion

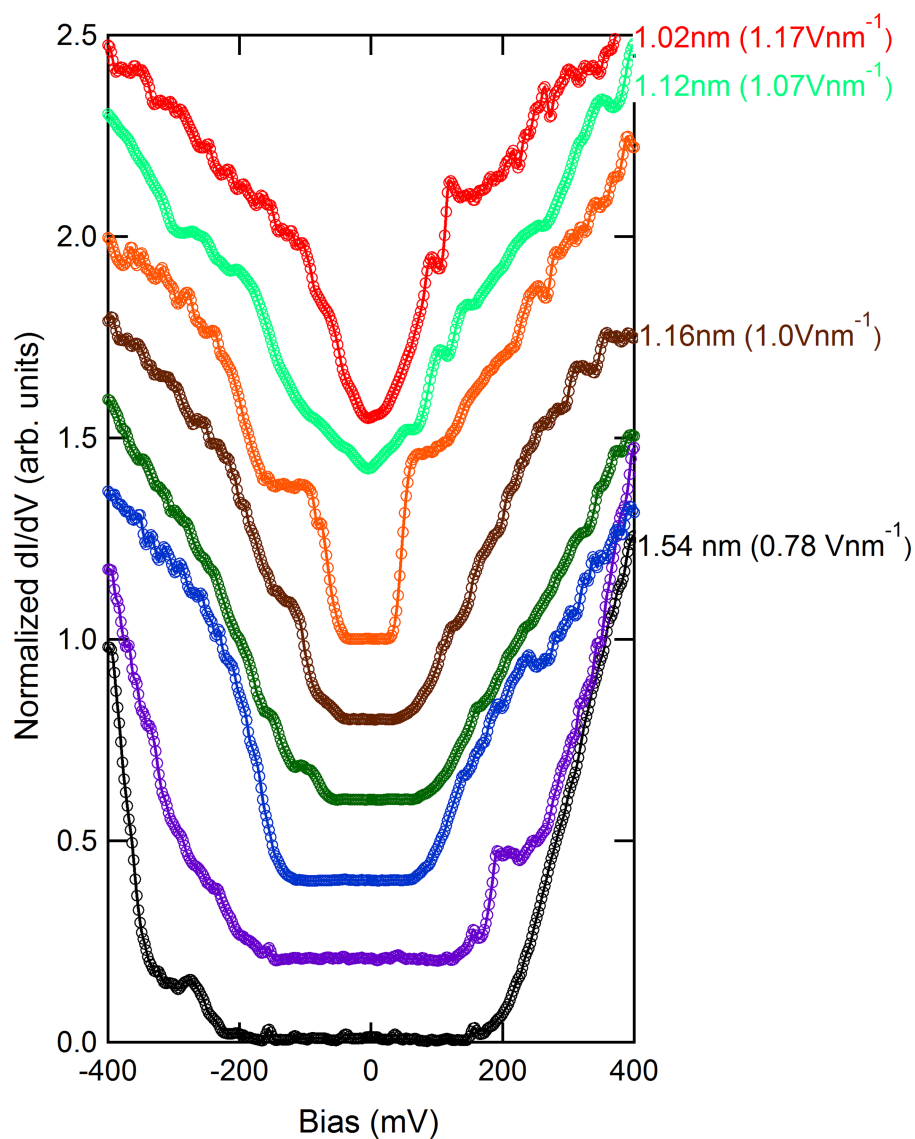
to be clearly resolved. Here we have used the ratio $(\Delta_p + \Delta_n)/\Delta_p \approx 1.4$ measured from the purple points to extrapolate Δ_n for undoped film. **b**, The valence edge (blue) and conduction edge (red) calculated from the fitted ARPES dispersion are shown for two high-symmetry directions: circles for K- Γ -K and squares for M- Γ -M.



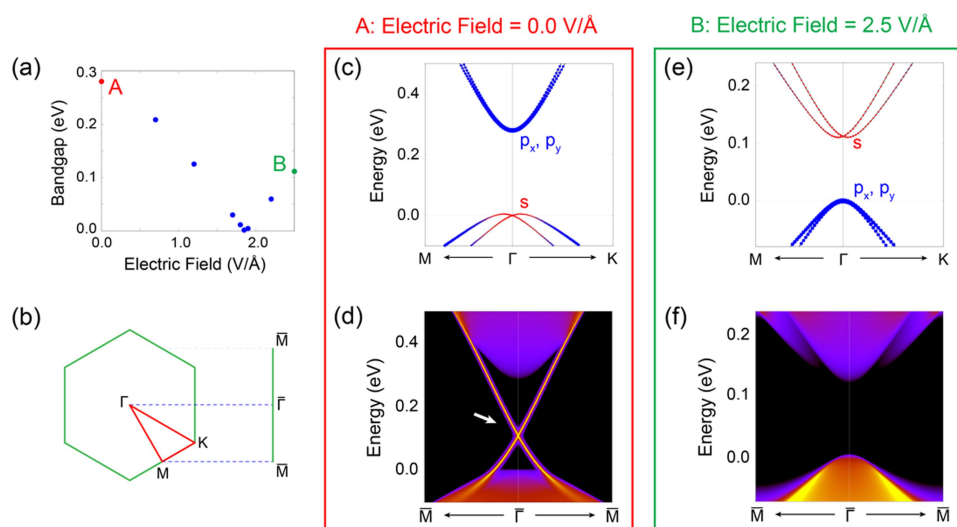
Extended Data Fig. 8 | Electric-field calculations from STM.

a, b, Tunnelling current as a function of relative tip-sample distance z for **a**, Au(111) (bias +500 mV), and **b**, thin film Na₃Bi (bias -300 mV). The top axis in **b** represents the total distance, s , between tip and sample determined as described in Methods section 'Calculating tip-sample separation and electric field'. The black lines in **a** and **b** are exponential fits. **c**, Energy-level schematic illustrating the effects of an image potential on the (purple trapezoidal) junction barrier, where s is the tip-sample

distance, and ϕ_{Sample} and ϕ_{Tip} are the work function of the sample and tip, respectively. Progression from blue, green, yellow, orange and red curves indicates modification of the apparent barrier height due to the imaging potential at decreasing tip-sample distances s . **d**, $\ln(I(s)/I_0)$ shown as a function of distance for a square barrier model (blue dashed line), for DFT-LDA (black full line), for the image potential model (black dashed-dotted), and for the local orbital basis model (red line). Data are taken from figure 10 of ref. ⁴¹.

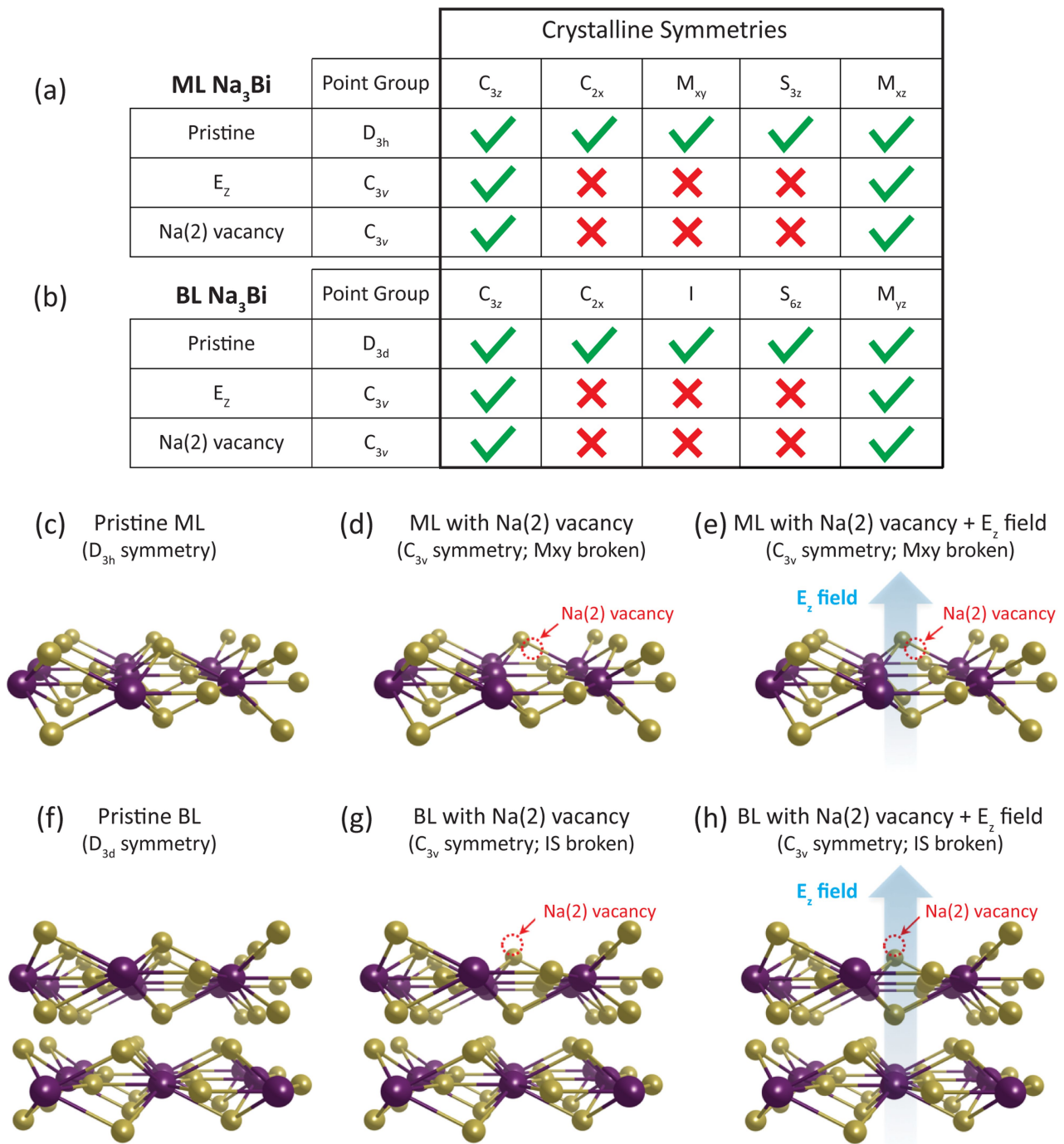


Extended Data Fig. 9 | Scanning tunnelling spectra of ML Na_3Bi . Shown are individual dI/dV spectra taken on ML Na_3Bi at different tip-sample separations (electric field) (shown at right). The spectra have been normalized and offset for clarity.



Extended Data Fig. 10 | DFT calculations of Na_3Bi layers with electric field. **a**, Calculated bandgap variation as a function of electric field for ML Na_3Bi with an Na(2) vacancy. The gap closes and reopens at about $1.85 \text{ V } \text{\AA}^{-1}$. **b**, 2D Brillouin zone and the projected 1D boundary Brillouin zone. **c–f**, Orbital-resolved band structures and edge spectra for ML Na_3Bi at electric fields of $0.0 \text{ V } \text{\AA}^{-1}$ (**c**, **d**) and $2.5 \text{ V } \text{\AA}^{-1}$ (**e**, **f**), which are marked by points A and B in **a**. **c**, Orbital-resolved band structures with SOC in the absence of electric field (the energy zero is set to be at the valence band maximum at the Γ point). The red dots represent the contribution from the Na s and Bi s atomic orbitals, and the blue dots

represent the contribution from the Bi p_x/p_y atomic orbitals. A band inversion is observed at the Γ point induced by SOC, indicating that ML Na_3Bi is a non-trivial 2D topological insulator. **d**, Projected edge spectrum (edge along the [010] direction) in the absence of electric field, showing topological edge states. Panels **e** and **f** show the corresponding results at an electric field of $2.5 \text{ V } \text{\AA}^{-1}$. In **e**, the band ordering at Γ is inverted compared to **c**, indicating a topological phase transition to a trivial insulator phase. This is confirmed by the disappearance of topological edge states as shown in **f**.



Extended Data Fig. 11 | Crystal structure and symmetries of ML and BL Na₃Bi. **a, b,** Tables showing the point group (PG) symmetries for ML and BL Na₃Bi, respectively. The second and third rows of each table indicate the effects of electric field (E_z) and an Na(2) vacancy, respectively. We note that at zero electric field both pristine and Na(2) vacancy ML and BL

Na₃Bi are 2D QSH insulators, with topological index $v \in \mathbb{Z}_2 = 1$. **c–h,** Crystal structure of Na₃Bi as a result of an Na(2) vacancy and electric field. **c, f,** Pristine ML (**c**) and BL (**f**) Na₃Bi; **d, g,** ML (**d**) and BL (**g**) Na₃Bi each with an Na surface vacancy; **e, h,** ML (**e**) and BL (**h**) Na₃Bi with Na surface vacancy plus electric field applied perpendicular to the sample.

Desymmetrization of cyclohexanes by site- and stereoselective C–H functionalization

Jiantao Fu¹, Zhi Ren¹, John Bacsá¹, Djamaladdin G. Musaev^{1,2} & Huw M. L. Davies^{1*}

Carbon–hydrogen (C–H) bonds have long been considered unreactive and are inert to traditional chemical reagents, yet new methods for the transformation of these bonds are continually being developed^{1–9}. However, it is challenging to achieve such transformations in a highly selective manner, especially if the C–H bonds are unactivated¹⁰ or not adjacent to a directing group^{11–13}. Catalyst-controlled site-selectivity—in which the inherent reactivities of the substrates¹⁴ can be overcome by choosing an appropriate catalyst—is an appealing concept, and substantial effort has been made towards catalyst-controlled C–H functionalization^{6,15–17}, in particular methylene C–H bond functionalization. However, although several new methods have targeted these bonds in cyclic alkanes, the selectivity has been relatively poor^{18–20}. Here we illustrate an additional level of sophistication in catalyst-controlled C–H functionalization, whereby unactivated cyclohexane derivatives can be desymmetrized in a highly site- and stereoselective manner through donor/acceptor carbene insertion. These studies demonstrate the potential of catalyst-controlled site-selectivity to govern which C–H bond will react, which could enable new strategies for the production of fine chemicals.

Selective reactions on certain cyclohexanes or polycyclic systems that contain appropriately positioned deactivating functionalities have been achieved^{18–22}. However, in the case of simple, electronically neutral cyclohexanes, good control of site- and stereoselectivity remains an unsolved challenge. Representative strategies in this area include carbene-induced C–H insertion²³, C–H oxidation¹⁸ and radical-induced C–H functionalization^{20,21,24}; however, their selectivities are limited. As such, the principal challenge of achieving these C–H functionalization processes in a highly site- and stereoselective manner has not yet been satisfactorily addressed.

Our group has previously reported the design and development of a series of chiral dirhodium catalysts with different steric environments (**1–4**; Fig. 1a, Extended Data Fig. 1), which are effective at catalysing C–H functionalization reactions of acyclic alkanes via donor/acceptor carbene insertion^{6,10,15,25}. We have also described the functionalization of cyclohexane using Rh₂(S-DOSP)₄ (**1**)¹⁰ (Fig. 1a). As the next challenge for our Rh(II)-catalysed C–H functionalization programme, and with the recent development of 2,2,2-trichloroethyl aryl diazoacetates as a more robust source of donor/acceptor carbenes²⁶, we became intrigued by the possibility of achieving site-selective C–H functionalization of more elaborate substrates such as substituted cyclohexanes. Here we describe the development and evaluation of a dirhodium catalyst, Rh₂(S-TPPTTL)₄ (**5**), leading to a site-selective carbene-insertion process with a high level of asymmetric induction. In particular, greater sophistication in stereocontrol is achieved, as the reaction generates three stereocentres in one step from an achiral substrate. For mono-substituted cyclohexanes, the catalyst is not only able to differentiate between C3 and C4, but also between C3 and C5, leading to desymmetrization of the substrate and generation of the products with high diastereoselectivity and enantioselectivity (Fig. 1).

We first examined *tert*-butylcyclohexane as our model substrate. With the bulky *tert*-butyl group in the preferred equatorial position,

there exist 11 different C–H bonds—excluding primary ones—that are electronically favoured towards C–H functionalization (Fig. 1b). The C1 axial position may be accessible for a structurally flexible catalyst, but is still largely unfavoured for steric reasons. In addition, C–H bonds at the C2 and C6 positions are likely to be too crowded for functionalization owing to the steric bulk of the rhodium carbene. For similar reasons, equatorial C–H bonds are more favoured than their axial counterparts. It is therefore reasonable to expect that three sites would be most favourable: the equatorial C–H bonds at C3, C4, and C5 (marked in red in Fig. 1b).

Initial exploratory studies of the reaction were conducted using 2,2,2-trichloroethyl 2-(4-bromophenyl)-2-diazoacetate (**7**) as the carbene source. When the relatively uncrowded catalyst Rh₂(S-DOSP)₄ (**1**) was used, the reaction gave primarily a mixture of three methylene-insertion products, **8–10**, with a small but noticeable quantity of the C1 methine-insertion product **11**. Catalysts **2–4** were sufficiently sterically hindered to block methine insertion, but still gave a mixture of products **8–10** (entries 2–4, Fig. 1c). During the course of these studies we evaluated a range of other established catalysts (see Supplementary Information 3 for the complete optimization study) as well as the new catalyst, Rh₂(S-TPPTTL)₄, which was readily prepared on a multi-gram scale in two steps. To our knowledge this is the first report of Rh₂(S-TPPTTL)₄, despite the fact that it is structurally related to Rh₂(S-TCPTAD)₄⁶ and other phthalimido-based catalysts that have been developed previously^{27–30}. In contrast to all the other catalysts we had studied, Rh₂(S-TPPTTL)₄ gave a very clean reaction (entry 5, Fig. 1c), favouring predominately a single methylene C–H functionalization product (**8**) with high site selectivity (>50:1 regio-meric ratio (r.r.)) and asymmetric induction (95% enantiomeric excess (e.e.)) (see Supplementary Information 7 for the X-ray structure of **8**). Notably, the product derived from C4 insertion (**10**) was not observed in the reaction catalysed by Rh₂(S-TPPTTL)₄. The products **8** and **9** are diastereomers and are formed through a desymmetrization event; this catalyst can therefore effectively distinguish between C3 and C4, and between the enantiotopic equatorial hydrogens at C3 and C5. To our knowledge, this has not been reported previously for any C–H functionalization of alkyl cyclohexanes.

Once we had established that Rh₂(S-TPPTTL)₄ is the optimal catalyst, we then sought to explore the reaction with other cycloalkanes (Fig. 2). Simple cycloalkanes were readily functionalized to produce **12–15** in good yield (72%–79%) and high enantioselectivity (90%–99% e.e.). With these benchmark data, we then explored a series of alkyl cyclohexanes to study the influence of the size of the substituent. All substrates underwent functionalization at the desired C3 position to form **16–22** with very high site selectivity, although in a few cases regioisomers were observed in minute amounts in the crude reaction mixture. Excellent levels of enantioselectivity (≥90% e.e.) were achieved for these substrates, indicating that the use of Rh₂(S-TPPTTL)₄ routinely results in high asymmetric induction at the carbene site. Particularly notable are compounds **18** and **19**, because regioisomers derived from possible reactions at the alkyl chains were formed only in trace amounts (>50:1 r.r.), even though these C–H bonds are very accessible. The

¹Department of Chemistry, Emory University, Atlanta, GA, USA. ²Cherry L. Emerson Center for Scientific Computation, Emory University, Atlanta, GA, USA. *e-mail: hmdavie@emory.edu

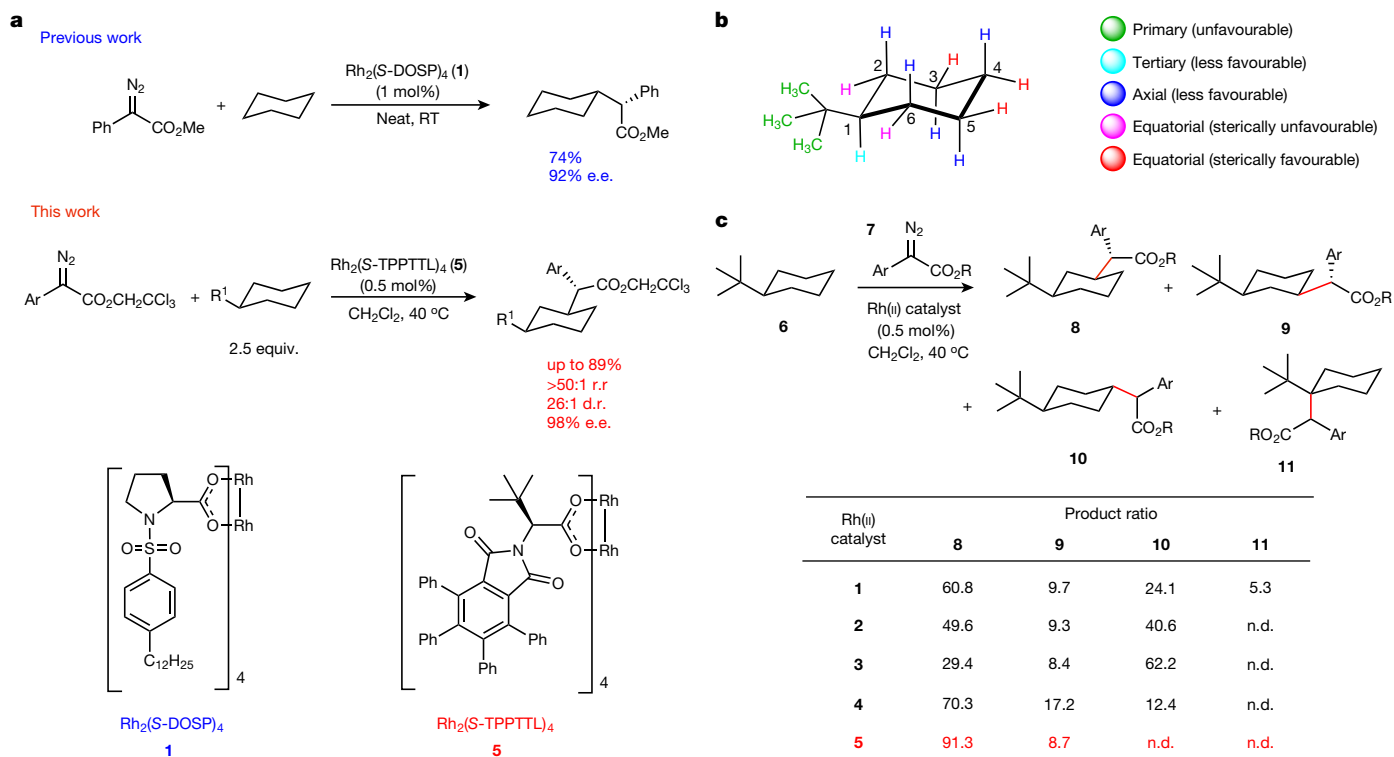


Fig. 1 | Background to the C–H functionalization of unactivated alkanes and its relationship to the current work. **a**, Functionalization of cyclohexanes with donor/acceptor carbenes. We have previously described the asymmetric C–H functionalization of cyclohexane using catalyst **1**. In this work we show that catalyst **5** is capable of functionalizing the C3 equatorial C–H bond of substituted cyclohexanes, leading to a stereoselective desymmetrization process. Ar, aryl or heteroaryl.

diastereomeric ratio (d.r.) was about 4:1 when the substituent was methyl or primary (**16–19**), but the diastereoselectivity improved steadily with secondary or tertiary substituents (**20–22**, 10–12:1 d.r.). These results indicate that the desymmetrization is more pronounced as the size of the alkyl substituent increases. Replacing the *tert*-butyl group in the model substrates with the trimethylsilyl (TMS) group led to a minimal change in reaction outcome (**23**). Furthermore, cyclohexanes bearing various ester groups also underwent clean functionalization to generate products **24–26** with good regio- and stereocontrol. The reaction was also compatible with various aryldiazoacetates, including some that contain heteroaryl donor groups (**27–38**). However, the level of asymmetric induction is sensitive to the steric bulk of the *para* substituent on the aryl group, as **33** and **34** were generated with high diastereoselectivity but lower enantioselectivity (47% and 79% e.e., respectively).

The selectivity of C–H functionalization by rhodium carbenes is generally considered to be governed by a combination of steric and electronic influences of the substrate and the catalyst of choice^{10,15,25}. Attempting to further evaluate these selectivity principles and to test the catalyst in more complex systems, we also subjected disubstituted alkyl cyclohexanes to the C–H functionalization reaction (Fig. 3). *cis*-1,2-Dimethylcyclohexane and *trans*-1,3-dimethylcyclohexane are of interest because they exist as 1:1 mixtures of enantiomeric chair forms. Both substrates underwent effective C–H functionalization, generating the products **39** and **42** with moderate to high levels of asymmetric induction (98% and 59% e.e.). However, the diastereoselectivity in the formation of **39** and **42** was quite low (2.2–3.7:1 d.r.), indicating that the reaction occurs with both enantiomeric chair forms as the substrate. *trans*-1,2-Dimethylcyclohexane is chiral and was reacted as the racemic mixture. Even so, the reaction was very effective, generating **40** with excellent site- and stereoselectivity. In addition, *trans*- and *cis*-1,4-dimethylcyclohexane are of interest as substrates because they

b, The structure of *tert*-butyl cyclohexane and the relative ease of functionalization at different positions. **c**, Optimization using the model substrate indicates that **5** can catalyse the functionalization of *tert*-butyl cyclohexane in a highly regioselective manner. See Supplementary Information 4 for experimental details and relevant spectra. Ar, (*p*-Br) C_6H_4 ; R, $\text{CO}_2\text{CH}_2\text{CCl}_3$.

enable us to evaluate the difference in reactivity between an axial and an equatorial C–H bond. Reaction with *cis*-1,4-dimethylcyclohexane resulted in C–H functionalization into a tertiary C–H site to form **43**. *trans*-1,4-Dimethylcyclohexane would be expected to exist primarily in the chair form with the two methyl groups in equatorial positions, yet this substrate was also capable of C–H functionalization to form **44**. However, the regioselectivity was lower (4.3:1 r.r.), presumably owing to unfavourable axial insertion and competition at other methylene sites. A substrate-competition experiment using an equal mixture of both 1,4-dimethylcyclohexane isomers indicated that an equatorial C–H bond reacted approximately 140 times faster than an axial C–H bond (see Supplementary Information 3 for experimental details). The equatorial preference observed here is much higher than that seen in other C–H functionalization reactions^{23,24}. Finally, the study was extended to *cis*- and *trans*-decalin; these substrates also gave clean transformations, forming **45** and **46** with excellent regio- and stereocontrol. The structure of **45** was confirmed by X-ray crystallography (see Supplementary Information 7).

To understand what makes $\text{Rh}_2(\text{S-TPPTTL})_4$ a successful catalyst in the desymmetrization of cyclohexanes, its structure was examined by X-ray crystallography and by DFT calculations (see Supplementary Information 6 and 7). The X-ray data indicate that the catalyst comprises a dirhodium core, with the four flanking phthalimido groups from the four S-TPPTTL ligands projecting upwards and adopting a ‘chiral crown’ shape²⁹; the structure is slightly distorted from perfect C_4 symmetry (Fig. 4) and is similar to that of other phthalimido dirhodium catalysts^{6,29}. A unique structural feature of $\text{Rh}_2(\text{S-TPPTTL})_4$ is the orientation of the 16 phenyl groups bound to the phthalimido ligands. The X-ray crystal structure shows that 12 of the phenyl groups are tilted to the right and four are tilted in the opposite direction. Further computational studies indicate that structure **47b** (the *M* configuration, in which all the phenyl groups are tilted to the right) is lower in energy

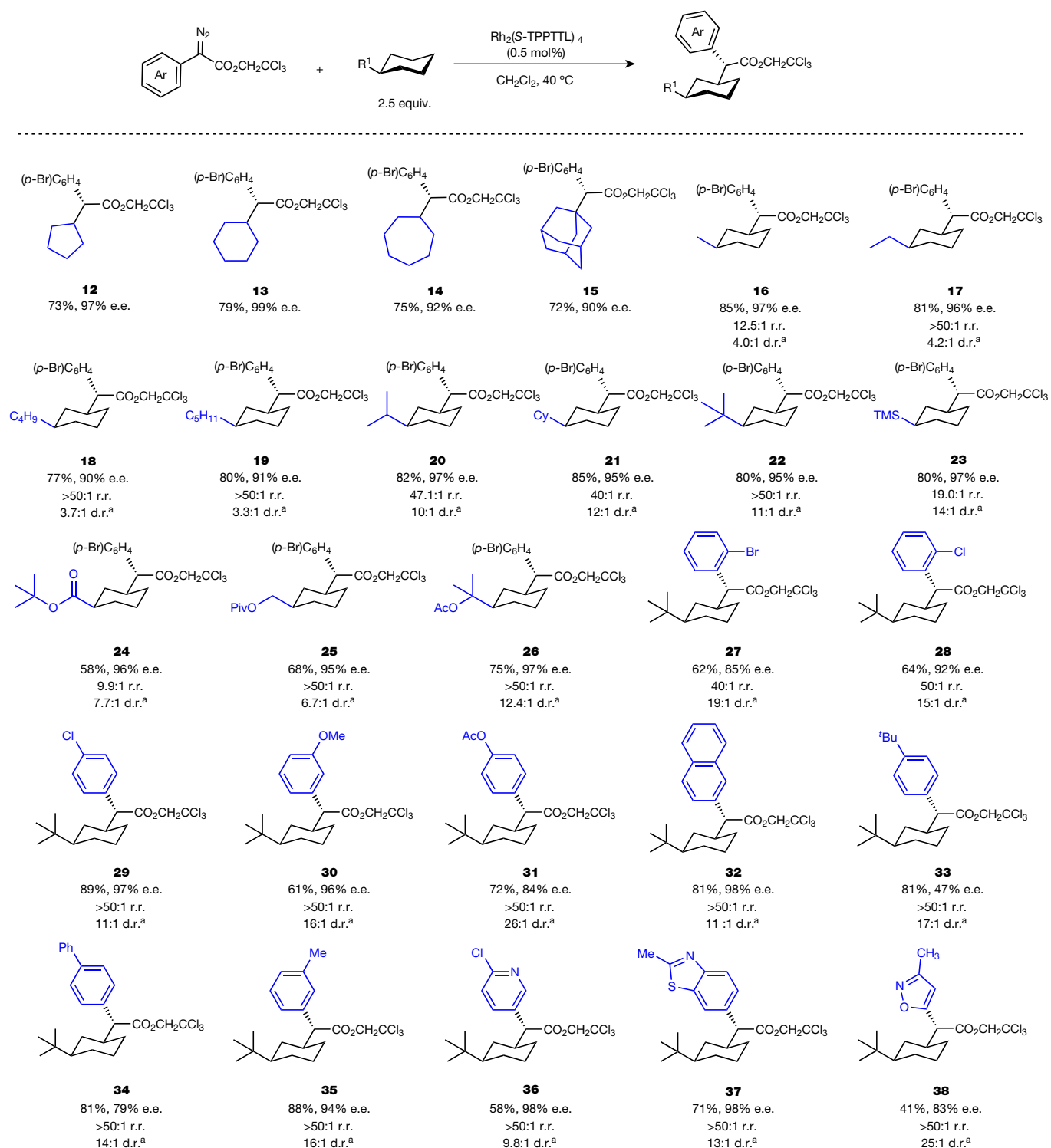


Fig. 2 | Scope of the reaction with respect to substrates and aryldiazoacetates. Simple cycloalkanes are readily functionalized with good yield and high enantioselectivity. For substituted cyclohexane substrates, high site-selectivity is routinely observed, resulting in C3 insertion via a desymmetrization event, although diastereoselectivity

is lower when the size of the substituent is small. The scope of aryldiazoacetates is broad, but sterically bulky *para*-substituents can lower enantioselectivity, as illustrated for compounds **33** and **34**. Heteroaryl donor groups are also compatible with this chemistry, as indicated by **36–38**. ^aNo ring diastereomers were observed.

than structure **47a** by 2.9 kcal mol^{−1}. Closer inspection of the structure reveals that the *tert*-butyl group of one ligand influences the tilt direction of the phenyl rings on the adjacent ligand. Indeed, attempts at calculating the energy of the complex in which all the phenyl groups are tilted to the left were not successful because the structure reverted back to the *M* configuration (see Supplementary Information 6 for details). Thus, the point chirality of the ligands induces a pseudo-*C*₄

propeller chirality in the complex by causing the 16 phenyl groups to tilt preferentially in one direction over the other. We propose that the orientations of these phenyl groups have a critical role in the observed selectivities.

Even though the ligands generate a deep pocket around the rhodium, computational studies indicate that binding of the carbene to this face (structure **48b**) is strongly preferred (by 15.5 kcal mol^{−1} as compared

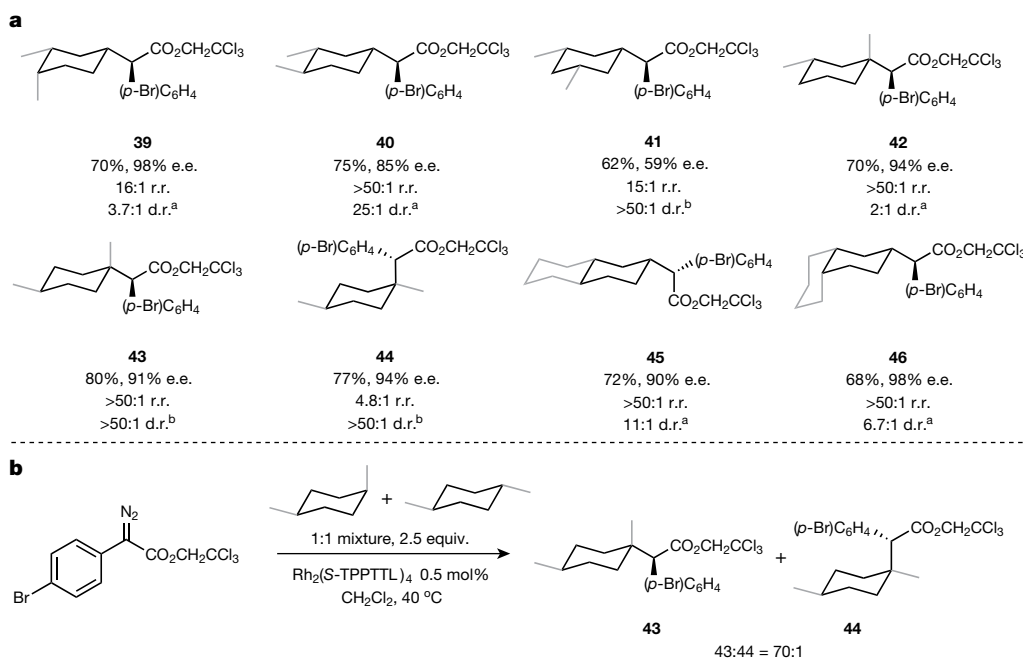


Fig. 3 | Functionalization of disubstituted cyclohexanes. **a**, C–H functionalization of disubstituted cyclohexanes is more challenging, and the selectivity is governed by the influence of the catalyst and the subtle electronic preferences of certain C–H bonds. Products are generally formed with high site- and stereoselectivity, although in a few cases the diastereomeric ratio is lower owing to reaction with both enantiomeric

chair forms. ^aNo ring diastereomers were observed. ^bOwing to symmetry, there are no side-chain diastereomers. **b**, A substrate competition study indicated that the equatorial C–H bond reacts 140 times faster than its axial counterpart, illustrating the general steric influence of the rhodium–carbene complex.

to structure **48a**), presumably because of the steric influence of the four *tert*-butyl groups. In structure **48b**, all of the phenyl groups are tilted in the same direction, and its other isomers—including the one with two oppositely tilted ligands—are higher in energy (see Supplementary Information 6 for details). Notably, comparing the structures of the free catalyst (**47b**) and the carbene-bound catalyst (**48b**) shows that the overall shape of the ligand framework changes to accommodate the carbene, which is indicative of an induced-fit model.

The next stage of the computational study aimed to understand how *tert*-butylcyclohexane approaches the rhodium carbene. We therefore calculated the structures and energies of several isomers of the Rh(carbene)(substrate) complexes³⁰. These studies reveal that attack of the rhodium carbene at the C4 position of the cyclohexane is very unfavourable, because the *tert*-butyl group would be pointing towards the ‘wall’ that is generated by the 16 phenyl groups of the catalyst. The most favourable structure of the Rh(carbene)(substrate) complex is **49**, in

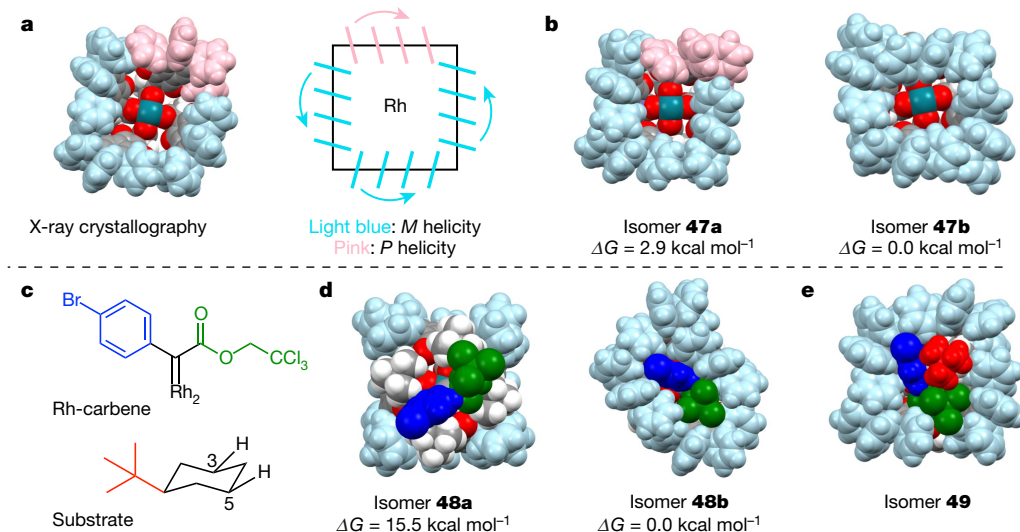


Fig. 4 | Rationalization of the observed selectivities. **a**, Top view of the crystal structure (left) and illustrative view of the helicity of the phenyl groups (right). **b**, Two calculated and energetically most stable isomers (**47a** and **47b**) of the catalyst. Isomer **47a** was optimized on the basis of X-ray data. **c**, Illustration of the colour-coding of atoms in the carbene and *tert*-butyl cyclohexane in the calculated structures: blue, donor group of carbene; green, acceptor group of carbene; red, *tert*-butyl group of the substrate. **d**, Calculated rhodium–carbene complexes. Energetically,

carbene binding to the top face (**48b**) is strongly favoured over binding to the bottom face (**48a**). **e**, Calculated lowest-energy Rh(carbene)(substrate) complex **49**. The atoms are colour-coded according to the default settings of the software Mercury: blue, rhodium; red, oxygen; white, hydrogen; grey, carbon. The highlighted atoms are coloured according to the direction in which the phenyl group is rotated: the phenyl groups in *M* helicity are coloured light blue, and those in *P* helicity are coloured pink.

which the *tert*-butyl group is pointing away from the ‘wall’ of the pocket and towards the opening of the binding face. This places one of the enantiotopic equatorial C3 hydrogens close to the carbene, which leads to a correct prediction of the asymmetric induction observed during desymmetrization. Examination of structure **49** shows that the shape of the catalyst adjusts once again to accommodate the substrate. Overall, these calculations show that Rh₂(S-TPPTTL)₄ has a high degree of flexibility to adjust its shape when the carbene and the substrate approach the catalytically active rhodium centre, which may explain why the reaction can be extended to disubstituted cyclohexanes and decalins.

In conclusion, we demonstrate that catalyst-controlled C–H functionalization of substituted cyclohexanes in a site- and stereoselective manner is a viable process. This study also further underscores the subtle controlling influences in the C–H functionalization reactions of donor/acceptor carbenes in the presence of appropriately designed dirhodium catalysts.

Data availability

Crystallographic data for the structures reported have been deposited at the Cambridge Crystallographic Data Centre, under deposition numbers CCDC 1855619, 1855620 and 1855295. Copies of the data can be obtained free of charge from www.ccdc.cam.ac.uk/data_request/cif. Complete experimental procedures and compound characterization data are available in the Supplementary Information, or from the corresponding author upon request.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0799-2>.

Received: 16 July 2018; Accepted: 25 October 2018;

Published online 19 December 2018.

- Yamaguchi, J., Yamaguchi, A. D. & Itami, K. C–H bond functionalization: emerging synthetic tools for natural products and pharmaceuticals. *Angew. Chem. Int. Ed.* **51**, 8960–9009 (2012).
- Davies, H. M. L. & Morton, D. Recent advances in C–H functionalization. *J. Org. Chem.* **81**, 343–350 (2016).
- Gutekunst, W. R. & Baran, P. S. C–H functionalization logic in total synthesis. *Chem. Soc. Rev.* **40**, 1976–1991 (2011).
- Wencel-Delord, J. & Glorius, F. C–H bond activation enables the rapid construction and late-stage diversification of functional molecules. *Nat. Chem.* **5**, 369–375 (2013).
- Fier, P. S. & Hartwig, J. F. Synthesis and late-stage functionalization of complex molecules through C–H fluorination and nucleophilic aromatic substitution. *J. Am. Chem. Soc.* **136**, 10139–10147 (2014).
- Liao, K. et al. Site-selective and stereoselective functionalization of non-activated tertiary C–H bonds. *Nature* **551**, 609–613 (2017).
- McMurray, L., O’Hara, F. & Gaunt, M. J. Recent developments in natural product synthesis using metal-catalysed C–H bond functionalisation. *Chem. Soc. Rev.* **40**, 1885–1898 (2011).
- Newhouse, T. & Baran, P. S. If C–H bonds could talk: selective C–H bond oxidation. *Angew. Chem. Int. Ed.* **50**, 3362–3374 (2011).
- Davies, H. M. L. & Manning, J. R. Catalytic C–H functionalization by metal carbenoid and nitrenoid insertion. *Nature* **451**, 417–424 (2008).
- Davies, H. M. L., Hansen, T. & Churchill, M. R. Catalytic asymmetric C–H activation of alkanes and tetrahydrofuran. *J. Am. Chem. Soc.* **122**, 3063–3070 (2000).
- He, J., Wasa, M., Chan, K. S. L., Shao, Q. & Yu, J. Q. Palladium-catalyzed transformations of alkyl C–H bonds. *Chem. Rev.* **117**, 8754–8786 (2017).
- Colby, D. A., Bergman, R. G. & Ellman, J. A. Rhodium-catalyzed C–C bond formation via heteroatom-directed C–H bond activation. *Chem. Rev.* **110**, 624–655 (2010).
- Hartwig, J. F. & Larsen, M. A. Undirected, homogeneous C–H bond functionalization: challenges and opportunities. *ACS Cent. Sci.* **2**, 281–292 (2016).
- Davies, H. M. L. & Morton, D. Guiding principles for site selective and stereoselective intermolecular C–H functionalization by donor/acceptor rhodium carbenes. *Chem. Soc. Rev.* **40**, 1857–1869 (2011).
- Liao, K., Negretti, S., Musaev, D. G., Bacsa, J. & Davies, H. M. L. Site-selective and stereoselective functionalization of unactivated C–H bonds. *Nature* **533**, 230–234 (2016).
- Qin, C. et al. D₂-symmetric dirhodium catalyst derived from a 1,2,2-triarylcyclopropanecarboxylate ligand: design, synthesis and application. *J. Am. Chem. Soc.* **133**, 19198–19204 (2011).
- Qin, C. & Davies, H. M. L. Role of sterically demanding chiral dirhodium catalysts in site-selective C–H functionalization of activated primary C–H bonds. *J. Am. Chem. Soc.* **136**, 9792–9796 (2014).
- Chen, M. S. & White, M. C. Combined effects on selectivity in Fe-catalyzed methylene oxidation. *Science* **327**, 566–571 (2010).
- Czaplyski, W. L., Na, C. G. & Alexanian, E. J. C–H Xanthylation: a synthetic platform for alkane functionalization. *J. Am. Chem. Soc.* **138**, 13854–13857 (2016).
- Schmidt, V. A., Quinn, R. K., Brusoe, A. T. & Alexanian, E. J. Site-selective aliphatic C–H bromination using *N*-bromoamides and visible light. *J. Am. Chem. Soc.* **136**, 14389–14392 (2014).
- Quinn, R. K. et al. Site-selective aliphatic C–H chlorination using *N*-chloroamides enables a synthesis of chlorolissoclimide. *J. Am. Chem. Soc.* **138**, 696–702 (2016).
- Wasa, M. et al. Ligand-enabled methylene C(sp³)–H bond activation with a Pd(ii) catalyst. *J. Am. Chem. Soc.* **134**, 18570–18572 (2012).
- Chen, K., Eschenmoser, A. & Baran, P. S. Strain release in C–H bond activation? *Angew. Chem. Int. Ed.* **48**, 9705–9708 (2009).
- Dondi, D. et al. Regio- and stereoselectivity in the decatungstate photocatalyzed alkylation of alkenes by alkylcyclohexanes. *Chem. Eur. J.* **15**, 7949–7957 (2009).
- Liao, K. et al. Design of catalysts for site-selective and enantioselective functionalization of non-activated primary C–H bonds. *Nat. Chem.* **10**, 1048–1055 (2018).
- Guptill, D. M. & Davies, H. M. L. 2,2,2-Trichloroethyl aryldiazoacetates as robust reagents for the enantioselective C–H functionalization of methyl ethers. *J. Am. Chem. Soc.* **136**, 17718–17721 (2014).
- Saito, H. et al. Enantio- and diastereoselective synthesis of *cis*-2-aryl-3-methoxycarbonyl-2,3-dihydrobenzofurans via the Rh(ii)-catalyzed C–H insertion process. *Org. Lett.* **4**, 3887–3890 (2002).
- Kitagaki, S. et al. Enantiocontrol in tandem carbonyl ylide formation and intermolecular 1,3-dipolar cycloaddition of α -diazo ketones mediated by chiral dirhodium(ii) carboxylate catalyst. *J. Am. Chem. Soc.* **121**, 1417–1418 (1999).
- DeAngelis, A. et al. The chiral crown conformation in paddlewheel complexes. *Chem. Commun.* **46**, 4541–4543 (2010).
- Nakamura, E., Yoshikai, N. & Yamanaka, M. Mechanism of C–H bond activation/C–C bond formation reaction between diazo compound and alkane catalyzed by dirhodium tetracarboxylate. *J. Am. Chem. Soc.* **124**, 7181–7192 (2002).

Acknowledgements Financial support was provided by the National Science Foundation (NSF) under the CCI Center for Selective C–H Functionalization (CHE-1700982). D.G.M. gratefully acknowledges NSF MRI-R2 grant (CHE-0958205) and the use of the resources of the Cherry Emerson Center for Scientific Computation. NMR and X-ray instrumentation used in this work was supported by the NSF (CHE-1531620 and CHE-1626172).

Reviewer information Nature thanks V. Gevorgyan and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.F. performed the synthetic experiments. Z.R. and D.G.M. conducted the computational studies. J.B. conducted the X-ray crystallographic studies. J.F. and H.M.L.D. designed and analysed the synthetic experiments and prepared the manuscript.

Competing interests The authors declare no competing interests.

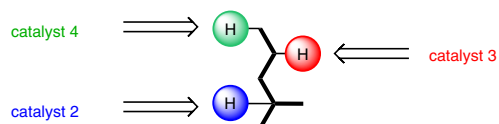
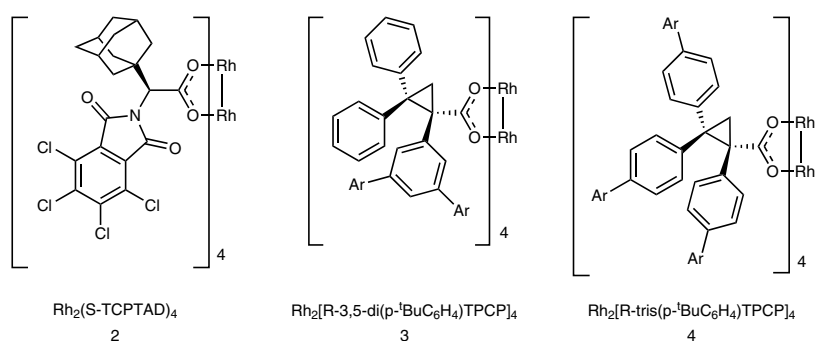
Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0799-2>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0799-2>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to H.M.L.D.
Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Extended Data Fig. 1 | Structures of previously established catalysts.
 We have previously shown that, through catalyst-directed C–H functionalization, the most accessible primary, secondary and tertiary

C–H bonds within a linear alkane substrate could be selectively functionalized by using catalyst **2**, **3** or **4**.

Origin of spatial variation in US East Coast sea-level trends during 1900–2017

Christopher G. Piecuch^{1*}, Peter Huybers², Carling C. Hay³, Andrew C. Kemp⁴, Christopher M. Little⁵, Jerry X. Mitrovica², Rui M. Ponte⁵ & Martin P. Tingley⁶

Identifying the causes of historical trends in relative sea level—the height of the sea surface relative to Earth’s crust—is a prerequisite for predicting future changes. Rates of change along the eastern coast of the USA (the US East Coast) during the past century were spatially variable, and relative sea level rose faster along the Mid-Atlantic Bight than along the South Atlantic Bight and the Gulf of Maine. Past studies suggest that Earth’s ongoing response to the last deglaciation^{1–5}, surface redistribution of ice and water^{5–9} and changes in ocean circulation^{9–13} contributed considerably to this large-scale spatial pattern. Here we analyse instrumental data^{14,15} and proxy reconstructions^{4,12} using probabilistic methods^{16–18} to show that vertical motions of Earth’s crust exerted the dominant control on regional spatial differences in relative sea-level trends along the US East Coast during 1900–2017, explaining most of the large-scale spatial variance. Rates of coastal subsidence caused by ongoing relaxation of the peripheral forebulge associated with the last deglaciation are strongest near North Carolina, Maryland and Virginia. Such structure indicates that Earth’s elastic lithosphere is thicker than has been assumed in other models^{19–22}. We also find a substantial coastal gradient in relative sea-level trends over this period that is unrelated to deglaciation and suggests contributions from twentieth-century redistribution of ice and water. Our results indicate that the majority of large-scale spatial variation in long-term rates of relative sea-level rise on the US East Coast is due to geological processes that will persist at similar rates for centuries.

Relative sea level (RSL) is the distance separating Earth’s crust from the sea surface. Changes in RSL can arise from any number of geological processes or climate dynamics that affect vertical land motion (VLM), sea surface height (SSH) or both. Identifying the processes responsible for RSL changes in historical coastal tide gauge records is important for anticipating future coastal hazards and constraining recent global-mean RSL rise^{6,7,22,23}.

The origin of large-scale spatial variation in centennial RSL trends as measured by tide gauges along the US East Coast has long been unclear^{1–13,24} (Extended Data Fig. 1); these trends are higher along the Mid-Atlantic Bight than along the South Atlantic Bight and the Gulf of Maine (Fig. 1a). Earlier studies argue that vertical crustal motions and gravity field changes tied to glacial isostatic adjustment (GIA)—Earth’s ongoing viscoelastic adjustment to the termination of the last ice age—are the dominant contributors to the spatial variation in RSL trends^{1,2}, such that higher trends on the Mid-Atlantic Bight reflect ongoing subsidence of the peripheral forebulge of the Laurentide Ice Sheet. Noting discrepancies between patterns of coastal RSL trends inferred from GIA models and tide gauge data, however, other work has highlighted the importance of ocean dynamics¹⁰, tectonic motions²⁴ or errors in GIA models³. More recently, investigations using updated GIA models, proxy reconstructions derived from saltmarsh sediment, and global positioning system (GPS) data hypothesize that, in addition to GIA, sediment compaction^{5,6}, dam retention^{7,8}, groundwater withdrawal^{7,8}, melting of the Greenland Ice Sheet⁹, ocean thermal expansion^{9,13} or

changes in ocean circulation^{10–12} contribute to the spatial variation in US East Coast RSL trends.

It is unclear whether these studies are contradictory. Formal error bars provided in these various studies do not account for important uncertainties inherent to the models, data and processes under consideration. Models of GIA suffer from uncertainties tied to ice history, mantle viscosity and lithospheric thickness³. Point-referenced tide gauge records, GPS data and saltmarsh-sediment proxy reconstructions can be short, sparse and fragmented; contaminated by local noise; and are seldom co-located alongside one another^{17,25} (for example, see Fig. 1a, b). Further complicating the interpretation of differences among studies is the existence of dependencies between models and data, and between different datasets, which have been ignored¹⁶. Rigorously determining the relative roles of VLM and SSH changes, and of GIA, in explaining the observed large-scale spatial structure in US East Coast RSL trends requires a mathematically coherent synthesis of available observations and models.

We use Bayesian data analysis^{16–18} to jointly infer the large-scale (larger than 500 km) spatial structure of centennial RSL trends on the US East Coast during 1900–2017. The contributions of VLM and SSH changes arising from GIA and other processes are determined at a common set of $0.5^\circ \times 0.5^\circ$ regularly spaced coastal grid points (see Methods). Inferences are based on 53 annual tide gauge RSL records¹⁴, VLM estimates from 42 GPS stations¹⁵, proxy RSL reconstructions derived from radiocarbon-dated sediment from 23 saltmarshes^{4,9,12} and 216 prior GIA model predictions based on three ice history models^{19–21} and 72 combinations of viscoelastic Earth structure parameters²³ (see Methods). RSL and other quantities of interest are modelled as processes with spatiotemporal dependencies that are described by uncertain parameters, including autocorrelation timescales, spatial ranges and error variances. Data are represented as noisy, biased and gappy versions of the underlying processes. We invert the model using Bayes’ rule to obtain the posterior probability distribution of the processes and parameters conditional on the data and prior estimates. The fully probabilistic solution provides rigorous uncertainty estimates, and allows for estimation of subtle pathwise statistics¹⁶, such as the probability density function associated with a spatially averaged value, the spatial variance in one process that is explained by another process, or whether a particular location features an extreme value.

Separating large-scale signals of interest from local processes and noise (see Methods; Figs. 1, 2), we find it very likely (probability $P = 0.98$) that the RSL trend averaged over the Mid-Atlantic Bight ($3.4 \pm 0.5 \text{ mm yr}^{-1}$) is larger than over the South Atlantic Bight ($2.7 \pm 0.6 \text{ mm yr}^{-1}$) and the Gulf of Maine ($2.2 \pm 0.7 \text{ mm yr}^{-1}$; Fig. 1c, f). All \pm ranges are 95% posterior credible intervals. The maximum RSL rise rate ($4.5 \pm 0.7 \text{ mm yr}^{-1}$) is likely ($P = 0.75$) to occur in North Carolina or Virginia, while the minimum trend ($1.3 \pm 0.8 \text{ mm yr}^{-1}$) is likely ($P = 0.86$) to occur in Florida or Maine (Fig. 1e). Similarly, it is likely ($P = 0.89$) that the average VLM rate over the Mid-Atlantic Bight ($-1.4 \pm 0.4 \text{ mm yr}^{-1}$) reflects stronger subsidence than along

¹Woods Hole Oceanographic Institution, Woods Hole, MA, USA. ²Harvard University, Cambridge, MA, USA. ³Boston College, Boston, MA, USA. ⁴Tufts University, Medford, MA, USA. ⁵Atmospheric and Environmental Research, Inc., Lexington, MA, USA. ⁶Los Gatos, CA, USA. *e-mail: cpiecuch@whoi.edu

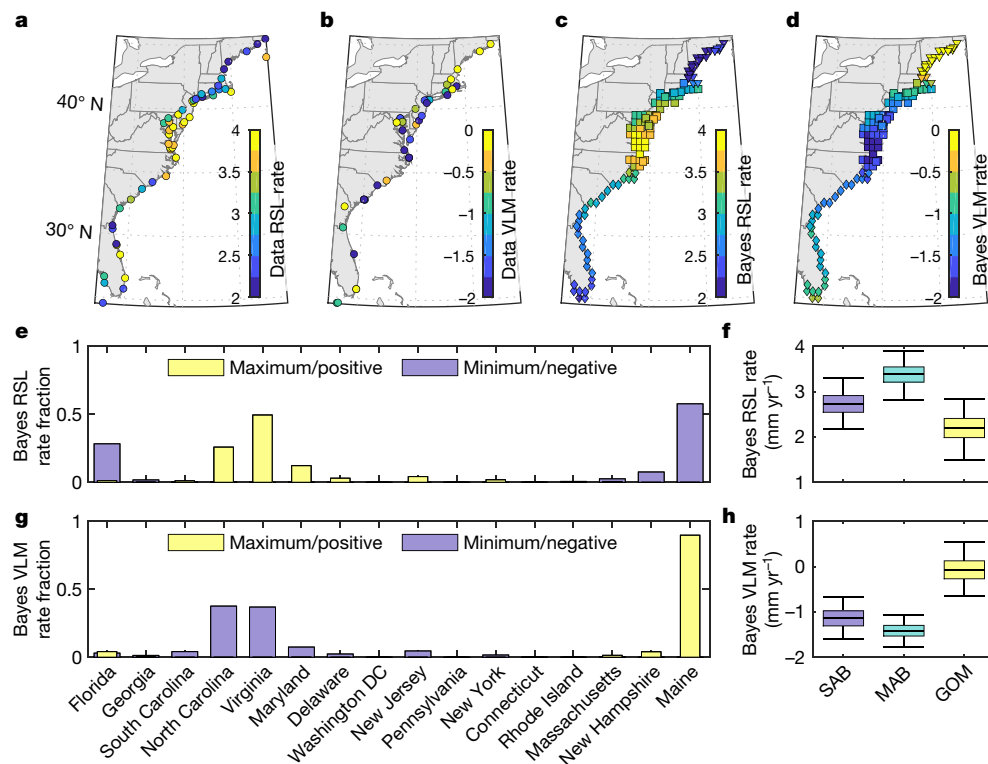


Fig. 1 | Rates of change. **a, b**, Trends in tide gauge RSL (**a**) and GPS station VLM (**b**). **c, d**, Median modelled RSL (**c**) and VLM (**d**) trends. Diamonds indicate the South Atlantic Bight (SAB), boxes indicate the Mid-Atlantic Bight (MAB) and triangles indicate the Gulf of Maine (GOM). **e, g**, Bayes rate extrema. Modelled probability that the maximum/most-positive or

minimum/most-negative RSL (**e**) and VLM (**g**) trend occurred in a given state. **f, h**, Model medians (lines), interquartile ranges (shading), and 95% credible intervals (whiskers) on SAB-, MAB- and GOM-averaged RSL (**f**) and VLM (**h**) trends.

the Gulf of Maine ($-0.1 \pm 0.6 \text{ mm yr}^{-1}$) and the South Atlantic Bight ($-1.1 \pm 0.5 \text{ mm yr}^{-1}$; Fig. 1d, h). We note that negative VLM reflects subsidence and hence contributes to sea-level rise. Correspondingly, the most negative VLM rate ($-2.5 \pm 0.6 \text{ mm yr}^{-1}$) is likely ($P = 0.75$) to

occur in the states that host the maximum sea-level rise, North Carolina or Virginia, whereas the most positive rate of VLM ($0.7 \pm 0.8 \text{ mm yr}^{-1}$) is very likely ($P = 0.90$) to occur in Maine (Fig. 1g). These regional spatial patterns are hinted at in the data (Fig. 1a, b), but the model

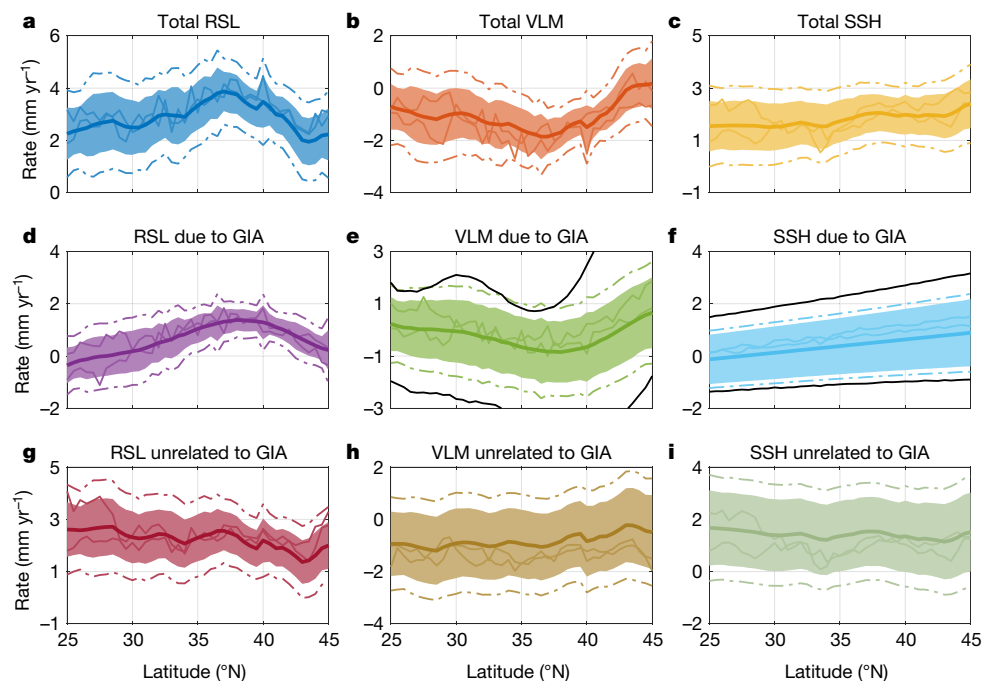


Fig. 2 | Latitudinal structure. **a–i**, Posterior median (thick line), 95% pointwise (light shade) and pathwise (dot-dashed) credible intervals, and two sample draws from the model solution (thin lines) for regional trends versus latitude for: **a**, RSL; **b**, VLM; **c**, SSH; **d**, GIA-driven RSL;

e, GIA-driven VLM; **f**, GIA-driven SSH; **g**, non-GIA RSL; **h**, non-GIA VLM; and **i**, non-GIA SSH. The 95% pathwise credible intervals are determined by broadening the 95% pointwise credible intervals until 95% of the solutions are encompassed. Black lines are prior 95% pointwise credible intervals.

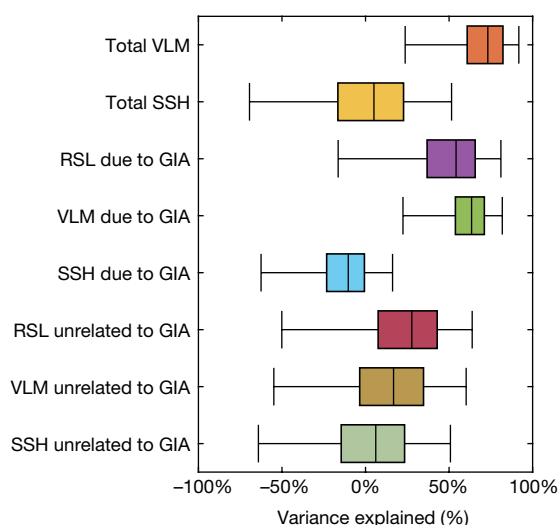


Fig. 3 | Contributions to spatial differences. Model median (black vertical lines), interquartile range (colour shading), and 95% credible interval (black whiskers) for the alongshore spatial variance in regional RSL linear trends during 1900–2017 explained by VLM or SSH related to GIA or other processes. Percentage variance V in x explained by y is defined as $100\% \times [1 - \text{var}(x - y)/\text{var}(x)]$, where var is the variance. Given the differences in sign convention (for example, a negative VLM rate corresponds to positive RSL trend), variances explained in RSL by VLM terms are computed by adding, rather than subtracting, the respective VLM component.

solutions are considerably smoother, owing to the suppression of noise associated with the spatiotemporal filtering and the joint assimilation of data streams involved in the Bayesian algorithm.

There is a striking visual correspondence between the latitudinal structures of the large-scale RSL and VLM trends (Fig. 2a, b). Adding posterior draws of regional RSL and VLM trends decreases the alongshore variance in the latter by a median of 73% (Fig. 3). Inferred regional SSH trends are comparatively more uniform (Fig. 2c). Nevertheless, there are hints of large-scale spatial structure, such that regional SSH trends are higher north of Cape Hatteras than south of that point (Fig. 2c).

More insight is gained by partitioning the regional trends (Fig. 2a–c) into GIA and other contributions (Fig. 2d–i; see Methods). We ascribe 69% of the large-scale variance in coastal VLM rates to GIA (median estimate). Estimated subsidence rates due to GIA are pronounced over the coastal Mid-Atlantic Bight, with the strongest trend ($-1.4 \pm 1.2 \text{ mm yr}^{-1}$) likely ($P = 0.81$) to be found in North Carolina, Maryland or Virginia, reflecting the collapse of the peripheral forebulge (Fig. 2e). Large-scale SSH trends due to GIA exhibit a statistically significant ($P > 0.99$) latitudinal gradient, with values increasing from south to north (Fig. 2f)—a consequence of geoid changes associated with mantle material flowing back to areas formerly overlain by the Laurentide Ice Sheet. The maximum RSL trend due to GIA ($2.0 \pm 0.4 \text{ mm yr}^{-1}$) is likely ($P = 0.69$) to be found in North Carolina, Maryland or Virginia, but is unlikely ($P = 0.23$) to be found in the states of Delaware or New Jersey, which are further north. This contrasts with past analyses of saltmarsh-sediment proxy reconstructions reasoning that the maximum rate of late-Holocene and ongoing RSL rise on the US East Coast due to relaxation of the peripheral forebulge is found in Delaware or New Jersey⁴. This apparent discrepancy arises from the uneven spatial distribution of the available saltmarsh reconstructions (see Supplementary Information).

Posterior GIA estimates are narrower than their corresponding priors (Fig. 2e, f), indicating that the posterior solutions are informative for distinguishing between the uncertain Earth structures and ice histories. Depending on ice and Earth-model choice, root-mean-square deviations between prior predictions and posterior solutions for RSL trends due to GIA are $0.4\text{--}1.9 \text{ mm yr}^{-1}$

(95% credible interval; see, for example, Fig. 4a). Viscosity ranges of $(0.3\text{--}0.5) \times 10^{21} \text{ Pa s}$ and $(2\text{--}3) \times 10^{21} \text{ Pa s}$ for the upper and lower mantle are, respectively, very likely ($P = 0.92$; Fig. 4b), and these are consistent with a recent study²⁶ comparing observed elevations of sea-level highstand markers along the US East Coast and the Caribbean to GIA model simulations during Marine Isotope Stages 5a and 5c. Our posterior GIA solutions are mostly constrained by the saltmarsh reconstructions and GIA priors, whereas the instrumental records have less influence (see Supplementary Information). The ice models adopted here^{19–21} were constructed by assuming an underlying viscoelastic Earth structure, and are not independent of the Earth models. Earth models assumed in earlier studies^{19–21} have viscosity structures similar to the prior Earth models favoured by our posterior solutions (Fig. 4a, b), so the viscosity ranges quoted above may be a natural consequence of the adopted ice models. However, the elastic lithospheric thickness favoured here (125 km; $P = 0.86$) is higher than in previous work^{19–22}. In keeping with physical intuition, the GIA solutions comprising our prior have a forebulge location that tends to be located further south with thicker lithosphere.

Given the long timescales characterizing GIA, the posterior solutions can be used to project RSL rise due to GIA into the future (Fig. 4c). RSL averaged over the South Atlantic Bight, the Mid-Atlantic Bight and the Gulf of Maine is predicted to rise by $2.5 \pm 3.7 \text{ cm}$, $10.3 \pm 2.2 \text{ cm}$ and $4.3 \pm 2.6 \text{ cm}$, respectively, during 2018–2100 owing to GIA (Fig. 4d). New York City and Washington DC are expected to experience respective increases of $9.6 \pm 4.5 \text{ cm}$ and $11.0 \pm 4.8 \text{ cm}$ (Supplementary Table 5), consistent with other recent estimates^{5,22,27}. Such changes related to inexorable geological processes will exacerbate predicted sea-level rise related to ocean thermal expansion, melting land ice and ocean circulation changes^{22,27}.

Although GIA is the first-order control on the regional spatial structure in centennial RSL trends, second-order contributions from other processes are evident in the posterior solution (Fig. 2g–i). RSL trends unrelated to GIA are very likely ($P = 0.95$) to increase from northern Maine to southern Florida (Fig. 2g). This structure is consistent with recent work²⁸ suggesting that ice melting, groundwater pumping and dam building globally since 1900 have caused higher RSL trends along the southern South Atlantic Bight than along the northern Gulf of Maine. Regional subsidence due to groundwater pumping and sediment compaction in South Carolina, North Carolina, and New Jersey reported previously^{5–7} does not feature strongly in our large-scale estimation, in that 95% credible intervals include zero (Fig. 2h), but residual analysis reveals significant local subsidence in these areas (see Supplementary Information). After removing the latitudinal trend, we find it very likely ($P = 0.91$) that Maine (at 43° N) is experiencing regional uplift (positive VLM) unrelated to GIA (Supplementary Table 6), corroborating a recent hypothesis⁸ that coastal Maine is uplifting isostatically in response to dam building over Québec, Canada. The coastal SSH expression of a poleward migration of the Gulf Stream during the twentieth century¹¹—higher trends on the Mid-Atlantic Bight than on the South Atlantic Bight and the Gulf of Maine—does not appear in our posterior solution (Fig. 2i; Supplementary Table 6). Furthermore, our solution is inconsistent with a dominant centennial contribution from ocean thermal expansion¹³ or declining Atlantic circulation and meridional heat transport^{29–31}, which would lead to higher RSL and SSH trends to the north of Cape Hatteras than to the south of that point³².

We identified the influences of VLM and SSH changes, arising from GIA and other processes, on the large-scale spatial variation in US East Coast RSL trends during 1900–2017. These findings clarify and build upon previous studies^{1–13,24}. Additional experiments demonstrate that our model solutions are robust to reasonable alternative selections for the priors on the scalar model parameters, study period duration and GPS dataset (see Supplementary Information). This work illustrates the value of jointly assimilating disparate data streams and modelling coupled physical processes within a coherent probabilistic framework for rigorous uncertainty quantification. In future work it will be useful

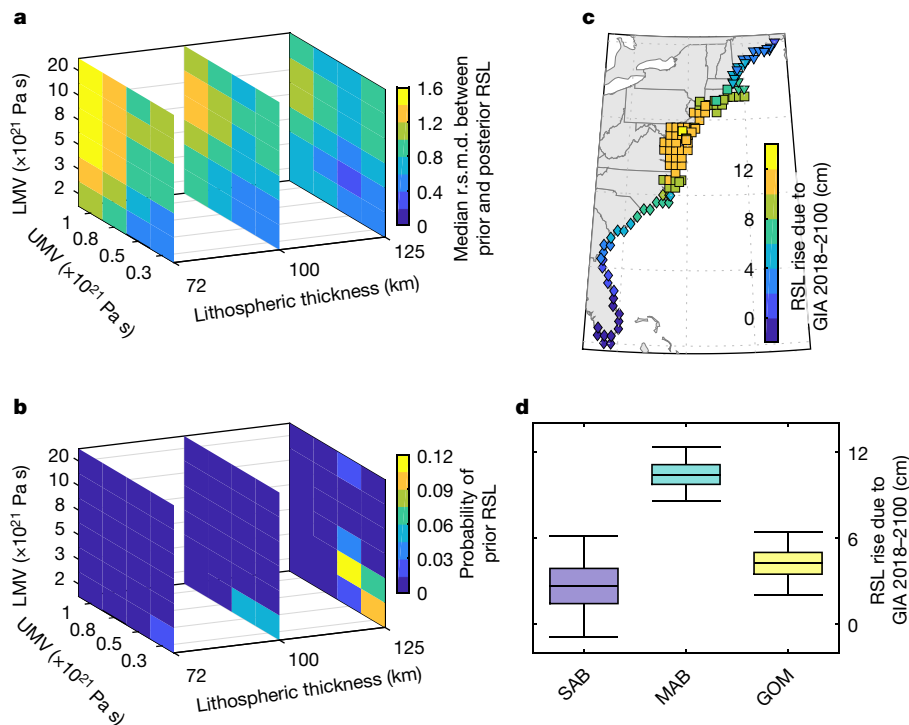


Fig. 4 | GIA-driven RSL trends. a, Median root-mean-square deviation between prior and posterior GIA-driven RSL trends as a function of rheological parameters used for the priors: lithospheric thickness, upper-mantle viscosity (UMV), and lower-mantle viscosity (LMV). **b**, Marginal posterior probability distribution for the probability that the best correspondence between prior and posterior solutions occurs for

a given combination of rheological parameters. **c**, Posterior medians of large-scale GIA-driven RSL change along the coast during 2018–2100. **d**, Posterior medians (lines), interquartile ranges (shading), and 95% credible intervals (whiskers) on the GIA-driven RSL rise during 2018–2100 averaged over the SAB, MAB and GOM.

to consider a broader region and incorporate spatial patterns associated with different mass sources to disaggregate terrestrial water storage and land ice contributions from large-scale RSL trends (Fig. 2g).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0787-6>.

Received: 22 January 2018; Accepted: 17 October 2018;

Published online 19 December 2018.

- Gornitz, V. & Seeber, L. Vertical crustal movements along the East Coast, North America, from historic and late Holocene sea level data. *Tectonophysics* **178**, 127–150 (1990).
- Peltier, W. R. & Tushingham, A. M. Influence of glacial isostatic adjustment on tide gauge measurements of secular sea level change. *J. Geophys. Res.* **96** (B4), 6779–6796 (1991).
- Davis, J. L. & Mitrovica, J. X. Glacial isostatic adjustment and the anomalous tide gauge record of eastern North America. *Nature* **379**, 331–333 (1996).
- Engelhart, S. E. & Horton, B. P. Holocene sea-level database for the Atlantic coast of the United States. *Quat. Sci. Rev.* **54**, 12–25 (2012).
- Kopp, R. E. Does the mid-Atlantic United States sea level acceleration hot spot reflect ocean dynamic variability? *Geophys. Res. Lett.* **40**, 3981–3985 (2013).
- Miller, K. G., Kopp, R. E., Horton, B. P., Browning, J. V. & Kemp, A. C. A geological perspective on sea-level rise and its impacts along the U.S. mid-Atlantic coast. *Earth. Fut.* **1**, 3–18 (2013).
- Karegar, M. A., Dixon, T. H. & Engelhart, S. E. Subsidence along the Atlantic Coast of North America: Insights from GPS and late Holocene relative sea-level data. *Geophys. Res. Lett.* **43**, 3126–3133 (2016).
- Karegar, M. A., T. H. Dixon, R. Malservici, J. Kusche, & S. E. Engelhart. Nuisance flooding and relative sea-level rise: the importance of present-day land motion. *Sci. Rep.* **7**, 1197 (2017).
- Engelhart, S. E., Horton, B. P., Douglas, B. C., Peltier, W. R. & Törnqvist, T. E. Spatial variability of late Holocene 20th century sea-level rise along the Atlantic coast of the United States. *Geology* **37**, 1115–1118 (2009).
- Douglas, B. C. Global sea level rise. *J. Geophys. Res.* **96** (C4), 6981–6992 (1991).
- Yin, J. & Goddard, P. B. Oceanic control of sea level rise patterns along the East Coast of the United States. *Geophys. Res. Lett.* **40**, 5514–5520 (2013).
- Kemp, A. C. et al. Late Holocene sea- and land-level change on the U. S. Southeastern Atlantic coast. *Mar. Geol.* **357**, 90–100 (2014).
- Wake, L., Milne, G. & Leuliette, E. 20th century sea-level change along the eastern US: unravelling the contributions from steric changes, Greenland Ice Sheet mass balance and Late Pleistocene glacial loading. *Earth Planet. Sci. Lett.* **250**, 572–580 (2006).
- Holgate, S. J. et al. New data systems and products at the Permanent Service For Mean Sea Level. *J. Coast. Res.* **29**, 493–504 (2013).
- Santamaría-Gómez, A. et al. Uncertainty of the 20th century sea-level rise due to vertical land motion errors. *Earth Planet. Sci. Lett.* **473**, 24–32 (2017).
- Tingley, M. P. & Huybers, P. Recent temperature extremes at high northern latitudes unprecedented in the past 600 years. *Nature* **496**, 201–205 (2013).
- Piecuch, C. G., Huybers, P. & Tingley, M. P. Comparison of full and empirical Bayes approaches for inferring sea-level changes from tide-gauge data. *J. Geophys. Res. Oceans* **122**, 2243–2258 (2017).
- Cressie, N. & Wikle, C. K. *Statistics for Spatio-Temporal Data* 1–588 (John Wiley & Sons, 2011).
- Peltier, W. R. Global glacial isostasy and the surface of the ice-age Earth: the ICE-5G (VM2) model and GRACE. *Annu. Rev. Earth Planet. Sci.* **32**, 111–149 (2004).
- Peltier, W. R., Argus, D. F. & Drummond, R. Space geodesy constrains ice age terminal deglaciation: the global ICE-6G_C (VM5a) model. *J. Geophys. Res. Solid Earth* **120**, 450–487 (2015).
- Lambeck, K., Rouby, H., Purcell, A., Sun, Y. & Sambridge, M. Sea level and global ice volumes from the Last Glacial Maximum to the Holocene. *Proc. Natl Acad. Sci. USA* **111**, 15296–15303 (2014).
- Love, R. E. The contribution of glacial isostatic adjustment to projections of sea-level change along the Atlantic and Gulf coasts of North America. *Earth. Fut.* **4**, 440–464 (2016).
- Hay, C. C., Morrow, E., Kopp, R. E. & Mitrovica, J. X. Probabilistic reanalysis of twentieth-century sea-level rise. *Nature* **517**, 481–484 (2015).
- Uchupi, E. & Aubrey, D. G. Suspect terranes in the North American margins and relative sea-levels. *J. Geol.* **96**, 79–90 (1988).
- Wöppelmann, G. & Marcos, M. Vertical land motion as a key to understanding sea level change and variability. *Rev. Geophys.* **54**, 64–92 (2016).
- Creveling, J. R., Mitrovica, J. X., Clark, P. U., Waelbroeck, C. & Pico, T. Predicted bounds on peak global mean sea level during marine isotope stages 5a and 5c. *Quat. Sci. Rev.* **163**, 193–208 (2017).
- Kopp, R. E. et al. Probabilistic 21st and 22nd century sea-level projections at a global network of tide-gauge sites. *Earth. Fut.* **2**, 383–406 (2014).
- Hamlington, B. D. et al. Observation-driven estimation of the spatial variability of 20th century sea level rise. *J. Geophys. Res. Oceans* **123**, 2129–2140 (2018).
- Rahmstorf, S. et al. Exceptional twentieth-century slowdown in Atlantic Ocean overturning circulation. *Nat. Clim. Chang.* **5**, 475–480 (2015).

30. Caesar, L., Rahmstorf, S., Robinson, A., Feulner, G. & Saba, V. Observed fingerprint of a weakening Atlantic Ocean overturning circulation. *Nature* **556**, 191–196 (2018).
31. Thornalley, D. J. R. et al. Anomalously weak Labrador Sea convection and Atlantic overturning during the past 150 years. *Nature* **556**, 227–230 (2018).
32. McCarthy, G. D., Haigh, I. D., Hirschi, J. J.-M., Grist, J. P. & Smeed, D. A. Ocean impact on decadal Atlantic climate variability revealed by sea-level observations. *Nature* **521**, 508–510 (2015).

Acknowledgements Funding came from Woods Hole Oceanographic Institution's Investment in Science Fund; Harvard University and from NSF awards 1558939, 1558966 and 1458921 and from NASA awards NNH16CT01C, NNX17AE17G and 80NSSC17K0698. We acknowledge conversations with S. Adhikari, B.D. Hamlington, F.W. Landerer, S.J. Lentz and P.R. Thompson.

Reviewer information *Nature* thanks M. King, R. Rietbroek and the other anonymous reviewer for their contribution to the peer review of this work.

Author contributions C.G.P. and P.H. jointly conceived the study. C.G.P., P.H. and M.P.T. formulated the model framework. C.C.H. and J.X.M. provided the GIA model solutions. A.C.K. provided the sea-level index points. C.G.P. performed the analyses and wrote the manuscript with input from all authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0787-6>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0787-6>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to C.G.P.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Observational data. We use data from 47 tide gauges on the US East Coast (Supplementary Table 1). Data span the South Atlantic Bight (south of Cape Hatteras), the Mid-Atlantic Bight (Cape Hatteras to Cape Cod) and the Gulf of Maine (north of Cape Cod). We also use six additional tide gauges along the southeastern Gulf of Mexico (Naples, Fort Myers, St Petersburg) and southwestern Atlantic Canada (Saint John, Yarmouth, Halifax) to better constrain the inference at the endpoints of the domain (southern South Atlantic Bight, northern Gulf of Maine). The annually averaged time series of mean RSL were downloaded from the Permanent Service for Mean Sea Level (PSMSL) Revised Local Reference database^{14,33} on 24 May 2018. Most records have at least 25 years of valid annual values. Exceptions include shorter records along the Florida coast (for example, Lake Worth Pier, Trident Pier and Daytona Beach), incorporated to fill a spatial gap in coverage. The dataset contains 3,248 gauge-years of data over 1900–2017 (around 52% completeness).

We also use vertical velocities and standard errors from 42 GPS stations on the US East Coast from the Université de La Rochelle 6a dataset¹⁵ (Supplementary Table 2). Stations feature 3–19 years of observations over 1995–2014 with more than 70% data completeness. Vertical velocities have been computed by researchers at the Université de La Rochelle, based on simultaneous fits of linear trends, position discontinuities, seasonal cycles and draconitic signals to daily station position estimates, while errors have been calculated using a power law plus white noise model for the residuals¹⁵. Values are expressed in the 2008 realization of the International Terrestrial Reference Frame³⁴. Standard errors provided with the dataset do not account for uncertainties associated with accurately realizing a stable International Terrestrial Reference Frame (such as those related to the origin and scale factor²⁵). Data were retrieved from Système d'Observation du Niveau des Eaux Littorales (SONEL) on 24 October 2017.

We also use proxy RSL reconstructions derived from radiocarbon-dated saltmarsh sediment, often called RSL index points, which are given as pairs of calibrated age and RSL (the difference between the altitude of a sample and the midpoint of its indicative range³⁵). We use 164 RSL index points from 23 saltmarshes culled from the Holocene database of Engelhart and Horton⁴ and updated to include data from northeastern Florida¹² (Supplementary Table 3). The formal uncertainties account for indicative range, radiocarbon dating, surveying and coring errors, but not for sediment consolidation errors. The geographic distribution of the data are highly uneven along the US East Coast. Because we wish to constrain contemporary trends related to GIA, we consider only RSL index points whose median calibrated age is between 2,000 and 150 years before present (where the 'present' is the year 1950). For a given saltmarsh site to be considered in the analysis, it must have at least three RSL index points with median ages within this specified range. We select this age range to predate a dominant anthropogenic influence on RSL, and to consider a period during which the contribution of GIA to RSL trends can reasonably be approximated as linear through time.

GIA model predictions. We incorporate predictions for contemporary VLM and SSH rates from 216 GIA models. Model predictions are distinguished by values used for lithospheric thickness (72 km, 100 km and 125 km), upper-mantle viscosity (0.3×10^{21} Pa s, 0.5×10^{21} Pa s, 0.8×10^{21} Pa s and 1.0×10^{21} Pa s), lower-mantle viscosity (2×10^{21} Pa s, 3×10^{21} Pa s, 5×10^{21} Pa s, 8×10^{21} Pa s, 10×10^{21} Pa s and 20×10^{21} Pa s), and ice history (ICE-5G¹⁹, ICE-6G²⁰ and ANU²¹). Model solutions are generated as described by ref. 23 and brought into the Bayesian framework as priors, as described below and in the Supplementary Information.

Bayesian framework. We develop a Bayesian algorithm for analysing tide gauge records, GPS data, RSL index points and GIA model predictions. The algorithm is a hierarchical dynamical spatiotemporal model¹⁸. The basic design follows ref. 17, who describe an algorithm for analysing tide gauge data on the North American northeast coast. Generalizations are made to analyse a larger region; to assimilate GPS data, RSL index points and GIA model solutions; and to separate the regional signals from local noise. A residual analysis justifying the model's form given the data follows in the Supplementary Information.

Process level. We wish to model RSL, VLM and SSH due to GIA and other processes. Given the nature of the data, our approach is to distinguish two periods, during which the controls on RSL changes are expected to be different. The first period is the modern era (since 1900), during which anthropogenic forcing affects centennial RSL rise, and instrumental data are available. For this period, during which observations are precisely dated, we seek to infer RSL process at all times and locations. The second period is a pre-industrial period (between 2,000 and 150 years before present), during which geological effects are expected to have a dominant control on longterm RSL trends, and RSL index points are available. For this period, during which the RSL index points have uncertain ages, we seek to infer the RSL process only at a subset of times and locations.

First, we consider the instrumental period. We model the spacetime evolution of the modern RSL process, $y_k = [y_{1,k}, \dots, y_{N,k}]^T$, for time steps $k \in \{1, \dots, K\}$ and locations $n \in \{1, \dots, N\}$ as a spatial field of linear temporal trends superimposed

on a first-order autoregressive [AR(1)] process driven by spatially correlated temporal innovations:

$$y_k - \mathbf{b}t_k = r(y_{k-1} - \mathbf{b}t_{k-1}) + \mathbf{e}_k \quad (1)$$

Here t_k is the time at step k , r is the AR(1) coefficient, \mathbf{b} is the spatial vector of temporal trends, and \mathbf{e}_k is the sequence of innovations. All model parameters are listed in Supplementary Table 4. The decision to model the detrended RSL residuals as an AR(1) process was motivated by ref. 36, which demonstrates that this assumption is justifiable for annual changes. Time steps are centred on zero, such that $\sum_{k=1}^K t_k = 0$. We model \mathbf{e}_k as a zero-mean, temporally independent and identically distributed (IID), spatially correlated vector, $\mathbf{e}_k \sim \mathcal{N}(\mathbf{0}_N, \Sigma)$, where \sim is read 'is distributed as', $\mathcal{N}(\mathbf{p}, q)$ is the multivariate normal vector distribution with mean \mathbf{p} and covariance q , $\mathbf{0}_X$ is the $X \times 1$ column vector of zeroes, and Σ is the $N \times N$ spatial covariance matrix given by:

$$\Sigma_{ij} = (c_{ij})\sigma^2 \exp(-\varphi|s_i - s_j|) \quad (2)$$

In equation (2), σ^2 is the partial sill³⁷, φ is the inverse range, and $|s_i - s_j|$ is the distance between locations s_i and s_j . Since RSL fluctuations north of Cape Hatteras are uncorrelated with RSL variations south of Cape Hatteras^{38–40}, the matrix element c_{ij} equals 1 if s_i and s_j are both either north or south of Cape Hatteras (approximately 35.25° N), and equals 0 otherwise.

We partition the field of RSL trends \mathbf{b} into SSH (\mathbf{w}) and VLM (\mathbf{u}) components,

$$\mathbf{b} = \mathbf{w} - \mathbf{u} \quad (3)$$

Rates of VLM and SSH are decomposed into contributions due to GIA (denoted by subscript g) and unrelated to GIA (denoted by superscript prime):

$$\mathbf{u} = \mathbf{u}_g + \mathbf{u}' \quad (4)$$

$$\mathbf{w} = \mathbf{w}_g + \mathbf{w}' \quad (4)$$

Trends in VLM and SSH unrelated to GIA— \mathbf{u}' and \mathbf{w}' —are represented as Gaussian random fields with spatial structure, $\mathbf{u}' \sim \mathcal{N}(\alpha \mathbf{1}_N, \Omega)$ and $\mathbf{w}' \sim \mathcal{N}(\mu \mathbf{1}_N, \Pi)$, where $\mathbf{1}_X$ is a $X \times 1$ column vector of ones:

$$\Omega_{ij} = \omega^2 \exp(-\rho|s_i - s_j|) \quad (5)$$

and

$$\Pi_{ij} = \pi^2 \exp(-\lambda|s_i - s_j|) \quad (6)$$

Here α and μ are spatial means, ω^2 and π^2 are partial sills, and ρ and λ are inverse ranges. Trends in VLM and SSH related to GIA— \mathbf{u}_g and \mathbf{w}_g —are assigned prior distributions based on the 216 GIA model predictions (see Supplementary Information). The set of vectors $\{\mathbf{b}, \mathbf{u}, \mathbf{w}, \mathbf{u}_g, \mathbf{w}_g, \mathbf{u}', \mathbf{w}'\}$ represents large-scale, long-period contributions to the trend fields.

The full VLM process \mathbf{v} is modelled as a Gaussian field, $\mathbf{v} \sim \mathcal{N}(\mathbf{u}, \varepsilon^2 \mathbf{I}_N)$, with mean vector equal to the spatially correlated large-scale VLM field \mathbf{u} , and a spatially uncorrelated covariance matrix. Here \mathbf{I}_X is the $X \times X$ identity matrix and ε^2 is a nugget effect³⁷ parameterizing the influence of local unresolved random processes. Thus, the local component of the VLM process is $\mathbf{v} - \mathbf{u}$.

Second, we consider the proxy era. We are interested in RSL at N_d spacetime points, corresponding to a subset $N_s \leq N$ of locations (N_d will be the number of RSL index points and N_s will be the number of saltmarshes). We model the spatiotemporal evolution of the pre-industrial RSL process, $\mathbf{Y} = [Y_1, \dots, Y_{N_d}]^T$, at times $T = [T_1, \dots, T_{N_d}]^T$, as a spatial field of linear temporal trends related to GIA superimposed on a random spacetime residual process:

$$\mathbf{Y} = [\sum_{i=1}^{N_d} \mathbf{e}_i \mathbf{e}_i^T G(\mathbf{w}_g - \mathbf{u}_g) \mathbf{e}_i^T] \mathbf{T} + \mathbf{D} \boldsymbol{\iota} + \mathbf{f} \quad (7)$$

Here $\boldsymbol{\iota}$ is a vector of site-specific intercepts, represented as a spatially uncorrelated normal random field, $\boldsymbol{\iota} \sim \mathcal{N}(\beta \mathbf{1}_{N_s}, \kappa^2 \mathbf{I}_{N_s})$, with mean β and variance κ^2 ; \mathbf{f} is a zero-mean, IID spacetime process, $\mathbf{f} \sim \mathcal{N}(\mathbf{0}_{N_d}, \varepsilon^2 \mathbf{I}_{N_d})$, with variance ε^2 ; and \mathbf{e}_i is the i th standard basis function of \mathbb{R}^{N_d} . The matrices G and D are selection matrices of ones and zeros, which isolate the GIA-driven RSL trend ($\mathbf{w}_g - \mathbf{u}_g$) and the intercept ($\boldsymbol{\iota}$), respectively, at the relevant target location. For example, G_{ij} equals one if element $i \in \{1, \dots, N_d\}$ of \mathbf{Y} corresponds to target location $j \in \{1, \dots, N\}$, and equals zero otherwise.

Unlike modern RSL y_k , pre-industrial RSL \mathbf{Y} is modelled without residual autocorrelation in time. This choice is motivated by the nature of the RSL index points. Recall that we choose to infer \mathbf{Y} only when and where RSL index points are available. This choice is made to speed up the algorithm. Index points at a particular

saltmarsh are widely separated in time, typically by decades or centuries. Given these wide separation timescales, it is reasonable to assume that temporal autocorrelation between residual RSL values (deviations from the longterm trend) is negligible. The reasonableness of this assumption is corroborated by the residual analysis in the Supplementary Information. Were a longer time period considered, such that the dominant behaviour is nonlinear, different choices would need to be made for modelling the pre-industrial RSL process.

Data level. Given data from tide gauges at $M_k \leq N$ locations at time step k , we represent the data, $\mathbf{z}_k = [z_{1,k}, \dots, z_{M_k,k}]^\top$, as gappy, noisy and biased versions of the RSL process:

$$\mathbf{z}_k = H_k \mathbf{y}_k + \mathbf{d}_k + F_k(\mathbf{a}_k + \ell) \quad (8)$$

Here \mathbf{d}_k is a random error sequence, cast as a temporally IID, spatially uncorrelated Gaussian field, $\mathbf{d}_k \sim \mathcal{N}(\mathbf{0}_{M_k}, \delta^2 \mathbf{I}_{M_k})$, where δ^2 is a variance parameter. The site-specific data offsets ℓ are modelled as a spatially uncorrelated normal random field, $\ell \sim \mathcal{N}(\nu \mathbf{I}_M, \tau^2 \mathbf{I}_M)$, with mean ν and variance τ^2 , where M is the total number of tide gauge sites ($N \geq M \geq M_k \forall k$). The data error trends \mathbf{a} are also represented as a Gaussian random field without spatial correlation, $\mathbf{a} \sim \mathcal{N}(\mathbf{0}_M, \gamma^2 \mathbf{I}_M)$, where γ^2 is a variance parameter. Matrices H_k and F_k are selection matrices that isolate the process, data bias and error trend vectors at the data sites at time step k .

Given GPS data at $L \leq N$ locations, we model the data, $\mathbf{x} = [x_1, \dots, x_L]^\top$, as gappy, noisy versions of the underlying VLM process, $\mathbf{x} \sim \mathcal{N}(E\mathbf{v}, \Delta)$. Here E is a selection matrix, which isolates the process at the observation sites, and Δ is an uncorrelated error covariance matrix, populated along the diagonal with error variances provided with the Université de La Rochelle 6a vertical velocity dataset. While Δ does not reflect uncertainties related to the realization of an International Terrestrial Reference Frame, the impact of such systematic GPS data issues on the Bayesian inference can be gleaned from sensitivity experiments discussed in the Supplementary Information. It is because Δ is specified a priori that the nugget effect ε^2 is identifiable. (We note that the data nugget effect ε^2 is distinct from the process variance parameter ε^2 .)

An important difference between tide gauge records and GPS data are that the former are spatiotemporal data (indexed in both space and time), whereas the latter are spatial data (indexed only in space). Whereas tide gauge records cover the period 1900–2017 (with at least one gauge returning data for each year of the epoch), GPS data span only the period 1995–2014, with many records covering only a fraction of that period. This poses a challenge from the perspective of inferring centennial rates of change. It is common to assume that VLM operates at steady rates over decades to centuries, and thus that GPS data are representative of much longer periods²⁵. While not strictly true, this assumption is a useful approximation; standard errors of about 0.5 mm yr⁻¹ are typical for 5-year GPS time series⁴¹. Our approach is to regard GPS data as a large-scale, long-period signal superimposed on small-scale, short-period noise. Our model is designed such that the signal is meant to be absorbed by the spatially structured field \mathbf{u} , whereas the noise is supposed to be captured by the spatially unstructured residual $\mathbf{v} - \mathbf{u}$. The underlying assumption is that large-scale, short-period and small-scale, long-period behaviours are negligible.

Given the RSL index points, we model the uncertain values of RSL, $\mathbf{Z} = [Z_1, \dots, Z_{N_d}]^\top$, and age, $\mathbf{S} = [S_1, \dots, S_{N_d}]^\top$, as noisy versions of the latent RSL values and their ages, $\mathbf{Z} \sim \mathcal{N}(\mathbf{Y}, \Gamma)$ and $\mathbf{S} \sim \mathcal{N}(\mathbf{T}, \Xi)$. Here Γ and Ξ are diagonal error covariance matrices, whose values are the formal error variances for the RSL and age estimates, respectively, provided with the Holocene RSL databases^{4,12}.

We select a set of $N = 211$ target locations, at which we make inference, to be the combined set of $M = 53$ tide gauge locations, $L = 42$ GPS stations, $N_s = 23$ saltmarshes, along with 93 regularly spaced $0.5^\circ \times 0.5^\circ$ grid points along the coast from southern Florida to northeastern Maine where no observations are present (Fig. 1a–d).

Prior level. To close the model, we place proper, mostly conjugate⁴² priors on the model parameters. Generally, these priors are selected to be diffuse, such that they have little influence on the posterior (see Supplementary Information). However, there are some exceptions that are important for understanding the results in the main text.

Given our interest in large-scale processes (variable ocean dynamics, melting of ice sheets, and so on), we condition the inference by constraining the inverse range parameters φ , ρ and λ in equations (2), (5) and (6) such that corresponding length scales characterizing the spatially correlated RSL innovations \mathbf{e}_k , and trends in VLM \mathbf{u}' and SSH \mathbf{w}' unrelated to GIA have a 95% prior probability of falling between roughly 500 km and 2,000 km. However, posterior solutions are robust to such details of prior selection; nearly identical posterior solutions for regional trend vectors are produced if wider or narrower priors are used on these parameters to condition the inference to focus on the large scales of interest to geology and climate (see Supplementary Information). Moreover, past authors note that

providing a prior sense of spatial scale on the inverse range is sometimes necessary to ensure convergence of the algorithm used to draw samples from the posterior distribution^{17,43–47}.

Given our particular interest in GIA, we place informative priors on the VLM and SSH trend vectors \mathbf{u}_g and \mathbf{w}_g related to GIA. Specifically, we place multivariate normal priors on these fields, with mean vectors and covariance matrices defined based on the 216 GIA model predictions. See the Supplementary Information for more details.

Drawing samples from the posterior distribution. Using Bayes' rule, the process and data level equations (1)–(8), and the priors, we assume that the posterior probability distribution of the process and parameters given the available data breaks down as

$$\begin{aligned} p(\mathbf{y}, \mathbf{Y}, \mathbf{T}, \Theta | \mathbf{x}, \mathbf{z}, \mathbf{Z}, \mathbf{S}) &\propto p(\mathbf{x}, \mathbf{z}, \mathbf{Z}, \mathbf{S} | \mathbf{y}, \mathbf{Y}, \mathbf{T}, \Theta) \times p(\mathbf{y}, \mathbf{Y}, \mathbf{T}, \Theta) \\ &= p(\mathbf{y}_0) \times p(\mathbf{r}) \times p(\sigma^2) \times p(\varphi) \times p(\mu) \times p(\pi^2) \times p(\lambda) \times p(\alpha) \\ &\times p(\omega^2) \times p(\rho) \times p(\varepsilon^2) \times p(\delta^2) \times p(\nu) \times p(\tau^2) \times p(\gamma^2) \\ &\times p(\beta) \times p(\kappa^2) \times p(\varepsilon^2) \times p(\mathbf{w}_g) \times p(\mathbf{u}_g) \\ &\times p(\mathbf{b} | \mathbf{u}, \mathbf{w}_g, \mu, \pi^2, \lambda) \times p(\mathbf{u} | \mathbf{u}_g, \alpha, \omega^2, \rho) \times p(\mathbf{v} | \mathbf{u}, \varepsilon^2) \times p(\ell | \nu, \tau^2) \\ &\times p(\mathbf{a} | \gamma^2) \times p(\mathbf{x} | \mathbf{v}) \times p(\mathbf{z} | \beta, \kappa^2) \times p(\mathbf{Y} | \mathbf{u}_g, \mathbf{w}_g, \mathbf{T}, \mathbf{u}, \varepsilon^2) \times p(\mathbf{Z} | \mathbf{Y}) \\ &\times p(\mathbf{S} | \mathbf{T}) \\ &\times \prod_{k=1}^K [p(\mathbf{z}_k | \mathbf{y}_k, \delta^2, \ell, \mathbf{a}) \times p(\mathbf{y}_k | \mathbf{y}_{k-1}, \mathbf{b}, \mathbf{r}, \sigma^2, \varphi)] \end{aligned} \quad (9)$$

Here p is probability density, $|$ is conditionality, \propto is proportionality, and $\Theta = \{\mathbf{b}, \mathbf{u}, \mathbf{w}, \dots\}$ is the set of all model parameters. Above, we assume that the data are conditionally independent given the process and the parameters.

We draw samples from the posterior distribution using Markov chain Monte Carlo (MCMC) methods similar to those in ref. 17. We evaluate full conditional distributions using a Gibbs sampler, with Metropolis steps used for the inverse range parameters. We perform 400,000 MCMC iterations, setting the initial process values to zero, and randomly drawing initial parameter values from the priors. We discard the first 200,000 draws to eliminate initialization transients, and keep only one out of every 200 samples to reduce the impacts of serial correlation between draws. Convergence is evaluated by comparing variance between and within chains. Results are based on three such 1,000-member chains concatenated together.

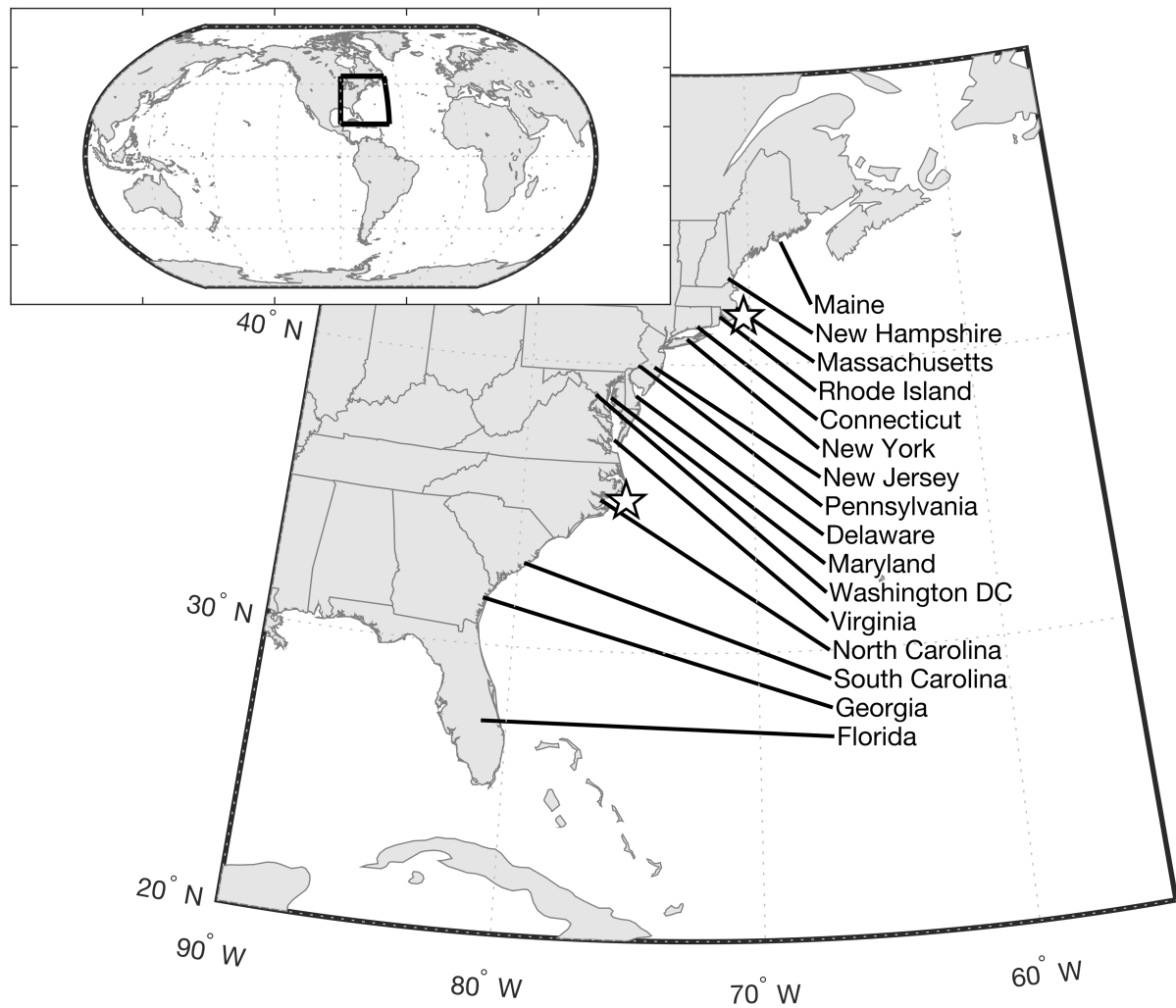
Code availability. The computer code used to run the Bayesian model and produce the results in this study, written in the MATLAB software environment, is available at the corresponding author's GitHub website (<https://github.com/christopherpiecuch>).

Data availability

The tide gauge and GPS data that support the findings of this study are available from the Permanent Service for Mean Sea Level (<http://www.psmsl.org/>) and Système d'Observation du Niveau des Eaux Littorales (<http://www.sonel.org/>), respectively. The proxy reconstructions are available from published databases^{4,12} and included with the model code (see 'Code availability' section). The GIA model predictions used to generate the results in this study are included with the model code (see 'Code availability' section). Maps in display items were produced using the Mapping Toolbox in MATLAB.

- Permanent Service for Mean Sea Level (PSMSL) *Tide Gauge Data* <http://www.psmsl.org/data/obtaining/> (PSMSL, 2018).
- Altamimi, Z., Collilieux, X. & Métivier, L. ITRF2008: an improved solution of the international terrestrial reference frame. *J. Geod.* **85**, 457–473 (2011).
- Engelhart, S. E., Horton, B. P. & Kemp, A. C. Holocene sea level changes along the United States' Atlantic Coast. *Oceanography* **24**, 70–79 (2011).
- Bos, M. S., Williams, S. D. P., Araújo, I. B. & Bastos, L. The effect of temporal correlated noise on the sea level rate and acceleration uncertainty. *Geophys. J. Int.* **196**, 1423–1430 (2014).
- Banerjee, S., Carlin, B. P. & Gelfand, A. E. *Hierarchical Modeling and Analysis for Spatial Data* 1–448 (Chapman and Hall, Boca Raton, 2004).
- Woodworth, P. L., Morales Maqueda, M. A., Roussenov, V. M., Williams, R. G. & Hughes, C. W. Mean sea-level variability along the northeast American Atlantic coast and the roles of the wind and the overturning circulation. *J. Geophys. Res. Oceans* **119**, 8916–8935 (2014).
- Thompson, P. R. & Mitchum, G. T. Coherent sea level variability on the North Atlantic western boundary. *J. Geophys. Res. Oceans* **119**, 5676–5689 (2014).
- Piecuch, C. G., Dangendorf, S., Ponte, R. M. & Marcos, M. Annual sea level changes on the North American Northeast Coast: influence of local winds and barotropic motions. *J. Clim.* **29**, 4801–4816 (2016).
- Santamaría-Gómez, A. & Mémin, A. Geodetic secular velocity errors due to interannual surface loading deformation. *Geophys. J. Int.* **202**, 763–767 (2015).

42. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian Data Analysis* 2nd edn, 1–668 (Chapman and Hall, Boca Raton, 2004).
43. Zhang, H. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Am. Stat. Assoc.* **99**, 250–261 (2004).
44. Tingley, M. P. & Huybers, P. A Bayesian algorithm for reconstructing climate anomalies in space and time. Part I: development and applications to paleoclimate reconstruction problems. *J. Clim.* **23**, 2759–2781 (2010).
45. Mannshardt, E., Craigmile, P. F. & Tingley, M. P. Statistical modeling of extreme value behavior in North American tree-ring density series. *Clim. Change* **117**, 843–858 (2013).
46. Tierney, J. E. & Tingley, M. P. A Bayesian, spatially-varying calibration model for the TEX₈₆ proxy. *Geochim. Cosmochim. Acta* **127**, 83–106 (2014).
47. Werner, J. P. & Tingley, M. P. Technical Note: Probabilistically constrained proxy age-depth models within a Bayesian hierarchical reconstruction model. *Clim. Past* **11**, 533–545 (2015).



Extended Data Fig. 1 | Study region. Map of the US East Coast and individual coastal states. Two white stars indicate Cape Cod (north) and Cape Hatteras (south), demarcating the three study regions: Gulf of Maine, Mid-Atlantic Bight and South Atlantic Bight.

Chemical differentiation, cold storage and remobilization of magma in the Earth's crust

M. D. Jackson^{1*}, J. Blundy² & R. S. J. Sparks²

The formation, storage and chemical differentiation of magma in the Earth's crust is of fundamental importance in igneous geology and volcanology. Recent data are challenging the high-melt-fraction 'magma chamber' paradigm that has underpinned models of crustal magmatism for over a century, suggesting instead that magma is normally stored in low-melt-fraction 'mush reservoirs'^{1–9}. A mush reservoir comprises a porous and permeable framework of closely packed crystals with melt present in the pore space^{1,10}. However, many common features of crustal magmatism have not yet been explained by either the 'chamber' or 'mush reservoir' concepts^{1,11}. Here we show that reactive melt flow is a critical, but hitherto neglected, process in crustal mush reservoirs, caused by buoyant melt percolating upwards through, and reacting with, the crystals¹⁰. Reactive melt flow in mush reservoirs produces the low-crystallinity, chemically differentiated (silicic) magmas that ascend to form shallower intrusions or erupt to the surface^{11–13}. These magmas can host much older crystals, stored at low and even sub-solidus temperatures, consistent with crystal chemistry data^{6–9}. Changes in local bulk composition caused by reactive melt flow, rather than large increases in temperature, produce the rapid increase in melt fraction that remobilizes these cool- or cold-stored crystals. Reactive flow can also produce bimodality in magma compositions sourced from mid- to lower-crustal reservoirs^{14,15}. Trace-element profiles generated by reactive flow are similar to those observed in a well studied reservoir now exposed at the surface¹⁶. We propose that magma storage and differentiation primarily occurs by reactive melt flow in long-lived mush reservoirs, rather than by the commonly invoked process of fractional crystallization in magma chambers¹⁴.

Magma reservoirs occur at several depths within the crust and typically grow incrementally through the intrusion of dykes or sills^{1,11,13,16,17}. High melt fractions must sometimes be present in these reservoirs to produce eruptible, low-crystallinity magmas^{1,7–9,13}. However, geophysical data suggest that reservoirs have low melt fraction even beneath active volcanoes^{2–5} and crystal chemistry data indicate that long-term magma storage occurs at low or even sub-solidus temperatures^{6–9}. High melt fractions are therefore ephemeral; yet geochemical models typically assume that differentiation occurs by crystal fractionation from low-crystallinity magmas^{11,14}. Moreover, geochronological data demonstrate that crustal magma reservoirs can be long-lived, spanning hundreds of thousands to millions of years^{17–21}. Existing models of crustal magma storage and differentiation cannot reconcile these conflicting observations.

We use numerical modelling to investigate the storage and chemical differentiation of magma in crustal reservoirs. The model describes repeated intrusion of mafic to intermediate sills into the mid- to lower crust^{12,13,16,21–23}, the associated transport of heat via conduction and advection and, in a key advance, mass transport via reactive flow of buoyant melt through the compacting crystal framework¹⁰. Transport of chemical components by the melt modifies the local bulk composition, and melt fraction changes in response to the chemical reactions that maintain local thermodynamic equilibrium. Phase behaviour is modelled using a two-component, eutectic phase diagram that, although

greatly simplified compared to natural systems, captures the critically important impact of bulk composition on melting behaviour and the complex nonlinear relationships between composition, melt fraction and permeability (see Methods)¹⁰. Melting relationships obtained from the phase diagram approximate common crustal igneous systems (Extended Data Fig. 1). The concentration of an incompatible trace element is also modelled assuming a constant partition coefficient.

Typical results are shown in Fig. 1 (see also Supplementary Video 1). In this example, 100-m-thick basalt (mafic) sills are intruded randomly over a depth range of 600 m, initially at 18 km depth and then around a depth that is controlled by the density contrast between intruding magma and host mush, reflecting the evolving reservoir composition and melt fraction (see Methods). We emplace 7.8 km of basalt in total, at an average rate of 5 mm yr^{–1} that is typical of crustal magmatic systems^{22–24}, into solid crust with an initial geotherm^{21–23} of 20 K km^{–1}. Our example was chosen to facilitate comparison with data from a well studied deep crustal section^{16,21}. The key findings are replicated over the depth range of 10–30 km typical of many crustal magma reservoirs and following intrusion of intermediate as well as mafic magma, using model parameters over a wide range that is reasonable for such systems (see Methods and Extended Data Table 1).

Initially, following each sill intrusion, the melt fraction rapidly falls to zero so there is no persistent magma reservoir (Supplementary Video 1 and Extended Data Fig. 2). This is the 'incubation phase' of the incipient magma reservoir, observed also in models that neglect reactive flow^{22,23}. However, in our model, chemical differentiation occurs within each intrusion before it solidifies, with more evolved melt (enriched in the incompatible trace element) accumulating at the top of the intrusion, and more refractory and depleted crystals accumulating at the base. The rapid increase in crystallinity traps the magma at the site of intrusion, but differentiation creates compositional contrasts that cause the intrusion depth to increase progressively (Supplementary Video 1 and Extended Data Fig. 3a).

The incubation phase ends when the melt fraction is greater than zero between successive sill intrusions, whereupon a magma reservoir has formed (Fig. 1a; Supplementary Video 1). Melt is now persistently present, but melt fraction remains low except for a brief period after each new intrusion (Extended Data Fig. 2b). The reservoir comprises a mush, rather than a high-melt-fraction magma chamber. Reactive flow now considerably modifies the predicted reservoir behaviour compared to previous models^{22,23}.

Buoyant melt migrates upwards through the mush, accumulating in the upper part of the reservoir because it cannot travel beyond the solidus isotherm where the melt fraction and permeability fall to zero (Supplementary Video 1). Melt composition evolves as it flows into, and reacts with, progressively cooler mush. Reactive flow reduces, or removes, early-formed compositional contrasts, such that the locally varying melt fraction controls the depth of later sill intrusions, which decreases as melt migrates upwards (see Methods). This is the 'growing phase' of the reservoir.

The growing phase ends when melt accumulates below the solidus isotherm to form a high-melt-fraction (typically >0.7) layer

¹Department of Earth Science and Engineering, Imperial College London, London, UK. ²School of Earth Sciences, University of Bristol, Bristol, UK. *e-mail: m.d.jackson@imperial.ac.uk

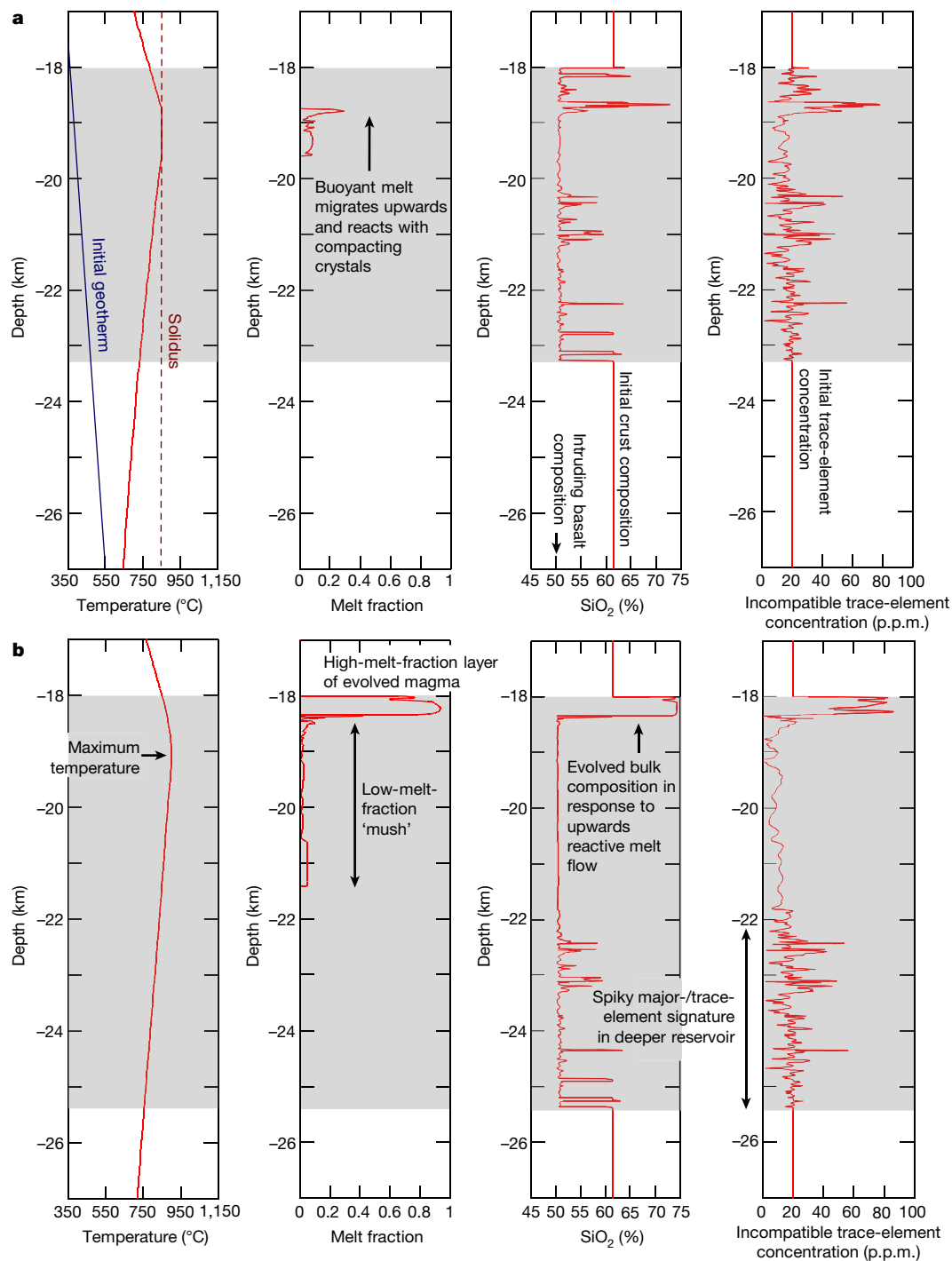


Fig. 1 | Temperature, melt fraction, bulk composition and incompatible trace-element concentration as a function of depth during the growing and active phases of the reservoir. a, b, Snapshots at 0.97 Myr (a) and 1.39 Myr (b) following the onset of sill intrusions, taken from Supplementary Video 1. At early times (not shown; see Extended Data Fig. 3a), during the incubation phase, individual sills cool rapidly. At later times (a), during the growing phase, a persistent mush reservoir forms but the melt fraction is low. Buoyant melt migrates upwards and begins to accumulate at the top of the mush. During the active phase (b), the accumulating melt forms a high-melt-fraction layer containing mobile

magma. The composition of the melt in the layer is evolved and enriched in incompatible trace elements. Elsewhere in the mush, the melt fraction remains low. At late times (not shown; see Extended Data Fig. 3b) during the waning phase, sill intrusions cease and the mush cools and solidifies. To illustrate the key processes, intruding basalt and crust are assumed in this example to have the same initial incompatible trace-element concentration. Shaded areas in all plots denote the vertical extent of basalt intrusions at that time. Equivalent results for sill intrusions at 10 km depth are shown in Extended Data Fig. 5.

overlying a thick (several kilometres), low-melt-fraction (typically <0.2) mush (Fig. 1b and Supplementary Video 1). The melt-rich layer contains chemically differentiated felsic magma and can grow to several hundreds of metres in thickness. Although not captured by the model, buoyant magma in the layer will be prone to leave the reservoir to produce shallower intrusions or volcanic eruptions^{25,26}.

Once magma leaves, a new layer grows by the same mechanism (see Methods).

This is the 'active phase', during which the reservoir can deliver evolved, low-crystallinity magma (Extended Data Fig. 2b). We suggest that, although geophysical surveys are probing active reservoirs, they image only the low-melt-fraction mush^{2–5}; the overlying

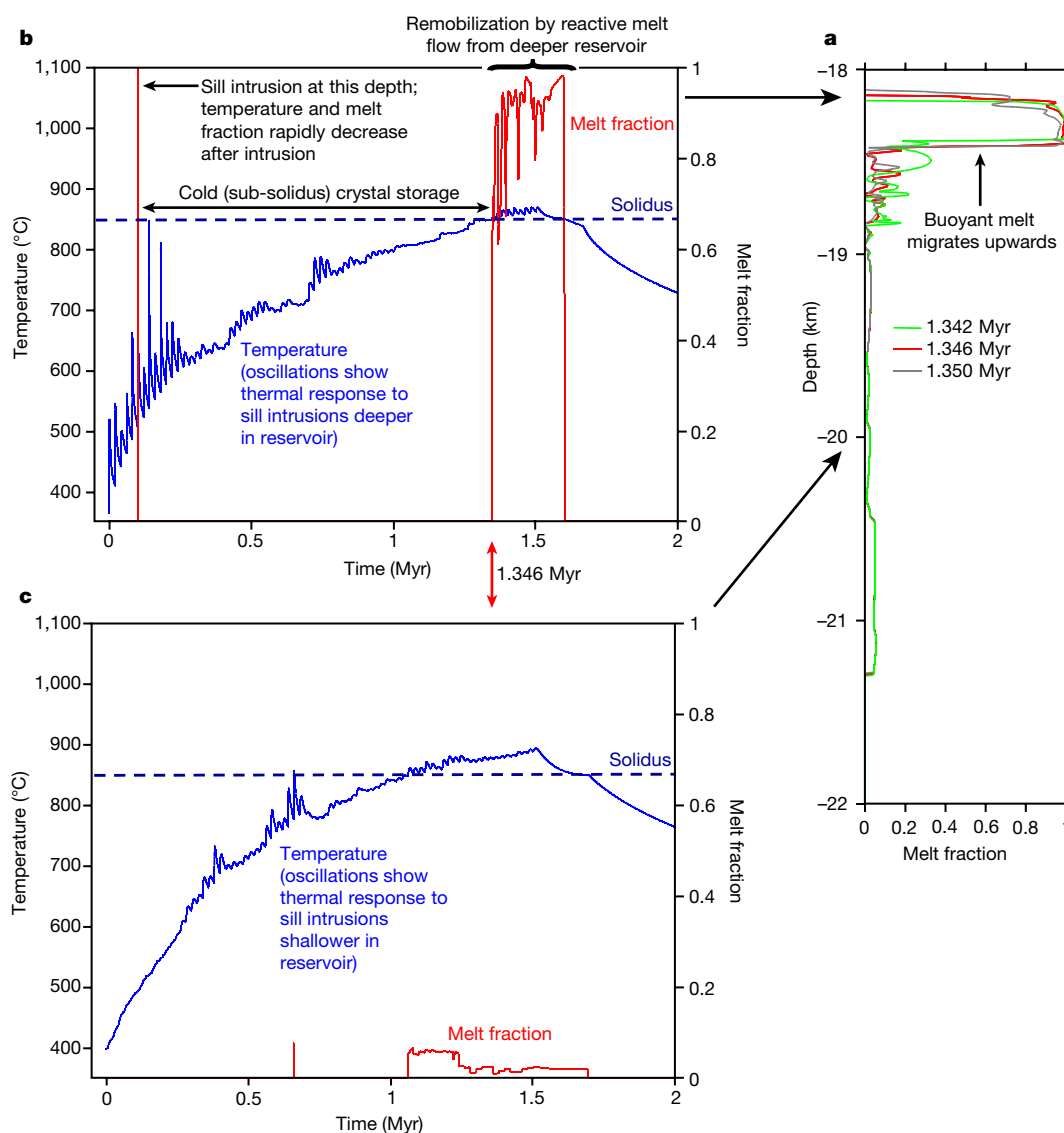


Fig. 2 | Cold storage and rapid remobilization of magma. **a**, Melt fraction as a function of depth at three different snapshots in time: at 1.346 Myr following the onset of sill intrusions and at 4 thousand years (kyr) before and 4 kyr after that time. Reactive flow of buoyant melt produces a high-melt-fraction layer that migrates upwards. **b**, Temperature and melt fraction as a function of time at a depth of 18.2 km, close to the top of the reservoir. Similar results are obtained over the depth range 18–18.5 km. Early sill intrusions rapidly cool and crystallize. The crystals are kept in ‘cold storage’ at sub-solidus temperature, but the temperature gradually increases in response to sill intrusions deeper in the reservoir. Soon (<0.3 kyr) after the temperature exceeds the solidus, the

high-melt-fraction layer arrives at this depth (the red arrow denotes the corresponding snapshot in **a**) and the reservoir is remobilized: the melt fraction increases rapidly to form a low-crystallinity magma. The melt fraction increases much more rapidly and to a higher value than would be possible by melting alone. **c**, Temperature and melt fraction as a function of time at a depth of 20 km. Similar results are obtained over the depth range 18.5–21.5 km. The melt fraction remains low because reactive flow has left a refractory residue at this depth. There is no remobilization, despite the increase in temperature. Data extracted from Supplementary Video 1. Equivalent results for intrusions at 10 km depth are shown in Extended Data Fig. 6.

high-melt-fraction layers are not observed, because they are ephemeral or too thin to be resolved. Geophysical detection of such a layer would suggest that magma mobilization (and possible eruption) was imminent⁷.

When intrusion of new sills ends, reactive flow continues wherever the temperature is above the solidus but, overall, the reservoir cools. This is the ‘waning phase’ (Supplementary Video 1; Extended Data Fig. 3b) that persists until the mush has completely solidified (Extended Data Fig. 2b). If exhumed, the resulting body of rock is termed a deep crustal section, of which there are several natural examples^{16,21}.

During the active phase, the high-melt-fraction layer forms towards the top of the reservoir where the temperature is low, rather than at the highest temperature (Fig. 1b). This counter-intuitive result is a consequence of reactive flow, whereby melt accumulation causes the local bulk composition to evolve towards the eutectic. Melt composition in

more chemically complex systems will evolve towards other low-variance states such as cotectics, peritectics or multiple-saturation points (see Methods), but the overall behaviour will be similar. A key finding here is that high-melt-fraction layers in crustal mush reservoirs can form in response to changes in bulk composition caused by reactive melt flow, rather than by large increases in temperature.

Magma in a high-melt-fraction layer contains about 10% crystals (Fig. 2a). These ‘antecrysts’ can long pre-date magma formation, because they derive from crystallization of early sills at the top of the reservoir. Once formed, the antecrysts are stored at near- or sub-solidus temperature (that is, ‘cool’ or ‘cold’; Fig. 2b). The local temperature gradually increases in response to ongoing intrusion of sills deeper in the reservoir and, eventually, exceeds the solidus. Soon afterwards, buoyant, evolved melt, migrating upwards through the pore space, accumulates around these older antecrysts, causing the local melt

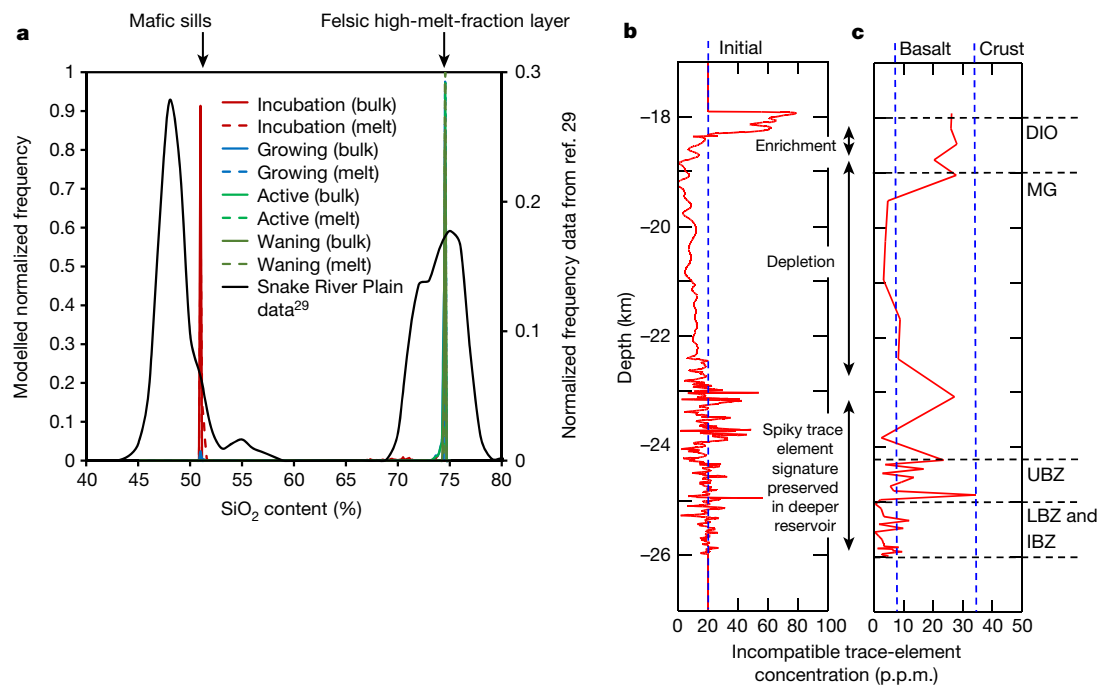


Fig. 3 | Geochemical consequences of reactive melt flow in crustal magma reservoirs. **a**, SiO_2 content of low-crystallinity (crystal fraction <30%) magmas. Solid curves show bulk magma composition (melt plus crystals); dashed curves show melt composition alone. The peak at low SiO_2 corresponds to magma within the intruding sills; the peak at high SiO_2 corresponds to magma within high-melt-fraction layers near the top of the reservoir. Also shown for comparison are data from the Snake River Plain²⁹. The bimodality is clear, although the basalt has a

lower SiO_2 content than modelled here. Results for different intruding sill compositions are shown in Extended Data Fig. 7. **b**, **c**, Modelled and observed neodymium concentration along a palaeo-vertical transect through the Upper Mafic Complex in the Ivrea-Verbano zone. LBZ, Lower Basal Zone; IBZ, Intermediate Basal Zone; UBZ, Upper Basal Zone; MG, main gabbro; and DIO, diorite¹⁶. Both modelled (**b**) and observed¹⁶ (**c**) data show a spiky profile at the base of the reservoir, depletion in the middle part of the reservoir and enrichment at the top.

fraction to increase rapidly and by far more than would be possible by heating alone (Fig. 2b)^{6,7,18,27}. Cold mush is remobilized here not by a substantial increase in temperature, but by buoyancy-driven reactive flow supplying evolved melt from deeper, more refractory parts of the reservoir, where temperature can be high but the melt fraction remain low (Fig. 2c). Remobilization is primarily caused by changes in local bulk composition, rather than temperature.

In our example, melt accumulation forms a low-crystallinity magma a few centuries after the local temperature exceeds the solidus, yet the

magma contains antecrysts formed up to about 1–1.4 millions of years (Myr) earlier (Fig. 2b). The range of antecryst ages reflects the timing of sill intrusions relative to the timing of melt accumulation. Crystal chemistry data show cool or cold storage and remobilization of older antecrysts hosted by younger felsic magma^{6–9}; our results suggest that this could result from reactive melt flow accumulating young, felsic melt within older mush. The antecrysts are not in equilibrium with the younger melt, creating disequilibrium crystal textures such as partial resorption. Flow of buoyant melt into the high-melt-fraction layer

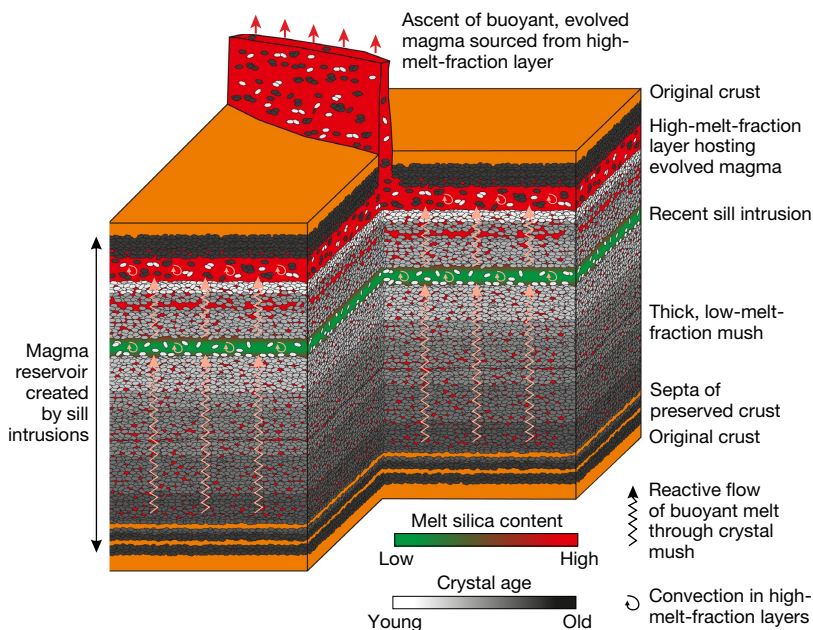


Fig. 4 | Reactive flow of buoyant melt at low melt fraction. Reactive flow is a critical mechanism controlling magma storage, accumulation and differentiation in mid- to lower-crustal reservoirs. The middle and lower parts of the reservoir comprise a thick (several kilometres) mush layer, with low and relatively uniform melt fraction, formed by early sill intrusions during the incubation and growing phases. This layer is typically imaged in geophysical data. During the active phase, the upper part of the reservoir comprises transient layers containing either intermediate/mafic or felsic magma, which can feed shallower intrusions or surface eruptions. The felsic magma layer is formed in response to changes in local bulk composition caused by upward reactive flow of buoyant melt through the mush. The evolved melt accumulates around older antecrysts, which may have formed during the earliest sill intrusions and hence long pre-date magma formation. In the schematic shown here, the felsic magma hosts a mixture of old and young antecrysts. The old antecrysts were formed during early sill intrusions; the young antecrysts formed during late sill intrusions at similar depth.

will drive convective overturn and homogenization before, or during, evacuation of magma, yielding a range of antecryst ages that may span the entire reservoir history²⁸.

Magmas in the high-melt-fraction layers have evolved composition. Conversely, magmas in the sills shortly after intrusion have compositions close to that of the intruded basalt. Low-crystallinity, mafic or felsic magmas can therefore leave the reservoir, but not magmas with intermediate composition. Many volcanic settings are characterized by bimodal volcanism (the ‘Daly Gap’), especially in oceanic settings (hotspots and island arc environments) and continental hotspots (Fig. 3a)^{14,15,29}. Our results suggest that compositional bimodality is another consequence of differentiation by reactive melt flow in mush reservoirs. However, not all systems show bimodality³⁰. Intermediate compositions could result from magma mixing¹⁵ or differentiation within multiple mush reservoirs comprising a vertically extensive magmatic system¹.

The modelled incompatible trace-element concentration in the solidified reservoir shows a characteristic pattern. Towards the base, the spiky signature produced by differentiation in each sill during the incubation phase is preserved (Fig. 3b). In the upper part, the profile is smoother and shows depletion relative to the initial concentration, reflecting extraction of melt. The top shows enrichment, reflecting accumulation of melt during the growing and active phases. Data from a deep crustal section show a similar pattern (Fig. 3c)¹⁶. We suggest that this pattern is another characteristic product of reactive melt flow in crustal mush reservoirs. Reactive melt flow at low melt fraction, rather than fractional crystallization at high melt fraction, is the dominant mechanism controlling magma storage, accumulation and chemical differentiation in the continental crust (Fig. 4).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0746-2>.

Received: 19 December 2017; Accepted: 2 October 2018;

Published online 3 December 2018.

- Cashman, K. V., Sparks, R. S. J. & Blundy, J. D. Vertically extensive and unstable magmatic systems: a unified view of igneous processes. *Science* <https://doi.org/10.1126/science.aag3055> (2017).
- Huang, H. H. et al. The Yellowstone magmatic system from the mantle plume to the upper crust. *Science* <https://doi.org/10.1126/science.aaa5648> (2015).
- Paulatto, M. et al. Magma chamber properties from integrated seismic tomography and thermal modeling at Montserrat. *Geochem. Geophys. Geosyst.* **13**, (2012).
- Hill, G. J. et al. Distribution of melt beneath Mount St. Helens and Mount Adams inferred from magnetotelluric data. *Nat. Geosci.* **2**, 785–789 (2009).
- Ward, K. M., Zandt, G., Beck, S. L., Christensen, D. H. & McFarlin, H. Seismic imaging of the magmatic underpinnings beneath the Altiplano-Puna volcanic complex from the joint inversion of surface wave dispersion and receiver functions. *Earth Planet. Sci. Lett.* **404**, 43–53 (2014).
- Rubin, A. E. et al. Rapid cooling and cold storage in a silicic magma reservoir recorded in individual crystals. *Science* <https://doi.org/10.1126/science.aam8720> (2017).
- Cooper, K. M. & Kent, A. J. R. Rapid remobilization of magmatic crystals kept in cold storage. *Nature* **506**, 480–483 (2014).
- Szymanowski, D. et al. Protracted near-solidus storage and pre-eruptive rejuvenation of large magma reservoirs. *Nat. Geosci.* **10**, 777–782 (2017).
- Andersen, N. L., Jicha, B. R., Singer, B. S. & Hildreth, W. Incremental heating of Bishop Tuff sanidine reveals preeruptive radiogenic Ar and rapid remobilization from cold storage. *Proc. Natl Acad. Sci. USA* **114**, 12407–12412 (2017).
- Solano, J. M. S., Jackson, M. D., Sparks, R. S. J. & Blundy, J. D. Evolution of major and trace element composition during melt migration through crystalline mush: implications for chemical differentiation in the crust. *Am. J. Sci.* **314**, 895–939 (2014).
- Glazner, A. F., Bartley, J. M., Coleman, D. S., Gray, W. & Taylor, R. Z. Are plutons assembled over millions of years by amalgamation from small magma chambers? *GSA Today* **14**, 4–12 (2004).
- Sisson, T. W., Ratajeski, K., Hankins, W. B. & Glazner, A. F. Voluminous granitic magmas from common basaltic sources. *Contrib. Mineral. Petrol.* **148**, 635–661 (2005).
- Rudnick, R. L. Making continental crust. *Nature* **378**, 571–578 (1995).
- Keller, B. C., Schoene, B., Barboni, M., Samperton, K. M. & Hesson, J. M. Volcanic–plutonic parity and the differentiation of the continental crust. *Nature* **523**, 301–307 (2015).
- Reubi, O. & Blundy, J. A dearth of intermediate melts at subduction zone volcanoes and the petrogenesis of arc andesites. *Nature* **461**, 1269–1273 (2009).
- Voshage, H. et al. Isotopic evidence from the Ivrea Zone for a hybrid lower crust formed by magmatic underplating. *Nature* **347**, 731–736 (1990).
- Coleman, D. S., Gray, W. & Glazner, A. F. Rethinking the emplacement and evolution of zoned plutons: geochronologic evidence for incremental assembly of the Tuolumne Intrusive Suite. *Calif. Geol.* **32**, 433–436 (2004).
- Barboni, M. et al. Warm storage for arc magmas. *Proc. Natl Acad. Sci. USA* **113**, 13959–13964 (2016).
- Deering, C. D. et al. Zircon record of the plutonic–volcanic connection and protracted rhyolite melt evolution. *Geology* **44**, 267–270 (2016).
- Frazer, R. E., Coleman, D. S. & Mills, R. D. Zircon U–Pb geochronology of the Mount Givens granodiorite: implications for the genesis of large volumes of eruptible magma. *J. Geophys. Res. Solid Earth* **119**, 2907–2924 (2014).
- Peressini, G., Quick, J. E., Sinigoi, S., Hofmann, A. W. & Fanning, M. Duration of a large mafic intrusion and heat transfer in the lower crust: a SHRIMP U–Pb zircon study in the Ivrea-Verbano zone (Western Alps, Italy). *J. Petrol.* **48**, 1185–1218 (2007).
- Annen, C., Blundy, J. D. & Sparks, R. S. J. The genesis of intermediate and silicic magmas in deep crustal hot zones. *J. Petrol.* **47**, 505–539 (2006).
- Karakas, O., Degruyter, W., Bachmann, O. & Dufek, J. Lifetime and size of shallow magma bodies controlled by crustal-scale magmatism. *Nat. Geosci.* **10**, 446–450 (2017).
- Crisp, J. A. Rates of magma emplacement and volcanic output. *J. Volcanol. Geotherm. Res.* **20**, 177–211 (1984).
- Malfait, W. J. et al. Supervolcano eruptions driven by melt buoyancy in large silicic magma chambers. *Nat. Geosci.* **7**, 122–125 (2014).
- Keller, T., May, D. A. & Kaus, B. J. P. Numerical modelling of magma dynamics coupled to tectonic deformation of lithosphere and crust. *Geophys. J. Int.* **195**, 1406–1442 (2013).
- Huber, C., Bachmann, O. & Dufek, J. The limitations of melting on the reactivation of silicic mushes. *J. Volcanol. Geotherm. Res.* **195**, 97–105 (2010).
- Bergantz, G. W., Schleicher, J. M. & Burgisser, A. Open-system dynamics and mixing in magma mushes. *Nat. Geosci.* **8**, 793–796 (2015).
- Ellis, B. S. et al. Rhyolitic volcanism of the central Snake River Plain: a review. *Bull. Volcanol.* **75**, 745 (2013).
- Bachmann, O. & Huber, C. Silicic magma reservoirs in the Earth’s crust. *Am. Mineral.* **101**, 2377–2404 (2016).

Acknowledgements M.D.J. and J.B. acknowledge funding from NERC Grant NE/P017452/1 “From arc magmas to ores (FAMOS): A mineral systems approach”. This paper is FAMOS contribution F05. M.D.J. also acknowledges sabbatical support from the Department of Earth Science and Engineering, Imperial College London, during which part of the research reported here was undertaken. R.S.J.S. acknowledges support from a Leverhulme Trust Emeritus Fellowship.

Reviewer information Nature thanks O. Bachmann, C. Till and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.D.J. wrote the code and produced the numerical results. J.B. prepared the phase equilibria model and calibrated this to experimental data. R.S.J.S. provided information on context and background for the study. All authors jointly designed the numerical experiments presented and drafted the manuscript text. M.D.J. prepared the figures.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0746-2>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0746-2>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.D.J.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Model formulation. To understand processes within crustal mush reservoirs, a quantitative model is required that includes three key features. First, the model must include the addition of hot magma or heat to initially solid crust, in order to create and grow the reservoir^{22,23,31–34}. Second, the model must include the relative motion of melt and crystals, to allow chemical differentiation^{10,34–37}. Third, the model must operate primarily at low melt fraction, consistent with a wealth of evidence that crustal magma reservoirs are normally low-melt-fraction mushes rather than high-melt-fraction magma chambers^{1–9,38,39}. At low melt fraction, a magma reservoir comprises a mush of crystals forming a solid framework with melt distributed along grain boundaries^{1,10,39–41}. At higher melt fraction, the reservoir comprises a slurry of melt containing suspended crystals that can flow via fractures, faults or other pathways to be intruded at shallower depths or erupt at the surface^{1,26,39,40}. The latter process is not modelled explicitly in this study.

The intrusion of magma to form sills can occur in numerous tectonic settings, providing both a source of heat and a source of magma that can differentiate to produce evolved melt^{1,13,16,17,21–23,31–33}. Here we follow earlier numerical approaches and model the repetitive intrusion of sills into the mid- to lower crust (modelling magma reservoirs at depths over the range 10–30 km), consistent with numerous contemporary magma reservoirs imaged in geophysical data and magma reservoirs interpreted to exist in deep crustal sections^{2–5,16,21–23,31–34,42–44}. It is assumed that the magma in the sills is delivered from some deeper reservoir in the crust or upper mantle. In most of the example cases shown, the intruding magma is mantle-derived basalt, recognizing that crustal magmatism is largely driven by basalt¹³ and consistent with numerous natural examples^{16,21,23,42–44}. However, in a later section we also show results for a case when the intruding sills contain more evolved (intermediate) magma.

Most models of repetitive sill intrusion do not include relative motion of melt and crystals and, therefore, there is no chemical differentiation: the bulk composition of the mush reservoir remains constant^{22,23,31–33}. Here, it is assumed that melt within the mush reservoir, produced by cooling and crystallization of the intruded sills and also heating and melting of the surrounding crust, is present along grain boundaries and forms an interconnected network at low melt fraction^{10,39–41}. The melt is buoyant because it is less dense than the surrounding crystals, so a pressure gradient is present that causes upward flow of melt relative to the crystalline matrix^{10,34,39}. The matrix can deform in response to melt flow^{45–47}. This coupled process of melt migration and matrix deformation is termed compaction⁴⁸. There is abundant evidence that compaction occurs in a wide variety of crustal igneous systems, and our assumptions are consistent with previous models of compaction^{10,26,34,48–52}.

Melt flow along grain boundaries in a mush reservoir allows efficient exchange of heat and mass between melt and solid phases, so that in most of the mush and over most of its lifetime, the phases remain in local thermal and chemical equilibrium⁵³. The local bulk composition of the mush therefore changes as the melt migrates upwards and the crystals compact downwards. To capture this, our model includes component transport and chemical reaction^{10,36,37,54}. The results shown here demonstrate that reactive flow of melt is a critically important process controlling the storage, accumulation and chemical evolution of magma within the mush reservoir.

Governing equations and method of solution. The governing equations and method of solution are modified from Solano et al.¹⁰. The enthalpy method is used to describe conservation of heat⁵⁵ and a binary eutectic phase diagram is used to describe solid and melt compositions, assuming local thermodynamic equilibrium^{10,37}.

In common with many previous studies, compaction is modelled using a modified version of the McKenzie formulation⁴⁸, assuming that deformation of the matrix occurs by melt-enhanced diffusion creep^{45–47}. This is reasonable in supra-solidus mush reservoirs deforming at low strain rates ($\ll 10^{-15} \text{ s}^{-1}$) and yields a Newtonian rheology for the mush⁵⁶. The matrix shear viscosity is assumed to be constant, but the matrix bulk viscosity has a power-law relationship with melt fraction^{26,50,57}. The melt is also assumed to have a Newtonian rheology, which is reasonable for crystal-free melts containing a few weight per cent water^{57,58}.

Surface tension and interphase pressure are neglected. The compaction formulation is currently being extended to include these potentially important effects, but a single, self-consistent model that includes phase change has not yet been presented^{59–63}. Differential stresses (imposed by tectonic forces) and magma chamber over-pressuring and loading⁶⁴ are also neglected, recognizing that at least some grain-boundary flow is essential to separate melt and crystals in a mush reservoir and buoyancy is always available to drive this. Volatiles are assumed to remain in solution, so are not present as a separate phase. In shallow crustal reservoirs, an evolved volatile phase probably plays an important part in controlling phase relations and melt flow, and in driving magma mobilization^{65,66}.

As outlined in Solano et al.¹⁰, the transport of heat, mass and components is modelled in one dimension, using a continuum formulation of the governing conservation equations. Typical sill intrusions and crustal mush reservoirs have high

aspect ratio^{2–5,16,21,38,42–44,67,68}. Given this, and the predominantly vertical flow of buoyant melt in the mush, a one-dimensional model is a reasonable starting point to determine the effects of reactive melt flow on magma storage and differentiation. However, a one-dimensional model does not admit the formation of high-porosity, sub-vertical channels caused by reactive infiltration instability⁵⁴. Numerical modelling in two dimensions has suggested that such channels are created during reactive melt flow in the mantle^{54,69,70}, but their formation and importance in crustal mush reservoirs is not yet clear. Future work should investigate whether additional controls on flow in crustal magma reservoirs are observed in two- and three-dimensional models. Such models are likely to be computationally expensive.

The Boussinesq approximation is applied, so density differences between solid and liquid are neglected except for terms involving gravity^{10,34,51}. Melt fraction and porosity are synonymous in this model. However, in contrast to previous models of crustal magma reservoirs, changes in local bulk composition resulting from melt migration mean that the local melt fraction here cannot be simply related via temperature to the melt fraction in the initial bulk composition (Extended Data Fig. 1c). This is a very important aspect of our model and one that pertains in both simple chemical systems (as employed here) and complex natural systems.

The governing equations can be expressed in dimensionless form as¹⁰

$$\frac{\partial h'}{\partial t'} = \kappa \frac{\partial^2 T'}{\partial z'^2} + \text{Ste} \frac{\partial}{\partial z'} ((1-\varphi) w'_s) \quad (1)$$

$$\frac{\partial C}{\partial t'} = -\frac{\partial}{\partial z'} ((1-\varphi) w'_s C_s) - \frac{\partial}{\partial z'} (\varphi w'_m C_m) \quad (2)$$

$$\frac{\partial I}{\partial t'} = -\frac{\partial}{\partial z'} ((1-\varphi) w'_s I_s) - \frac{\partial}{\partial z'} (\varphi w'_m I_m) \quad (3)$$

$$\frac{\partial}{\partial z'} \left(\varphi^{-\beta} \frac{\partial w'_s}{\partial z'} \right) = \frac{\mu' w'_s}{\varphi^\alpha} + (1-\varphi) \Delta \rho' \quad (4)$$

$$\varphi w'_m = -(1-\varphi) w'_s \quad (5)$$

where h is the enthalpy per unit mass; T is the temperature; t is the time; z is the vertical coordinate; w is the velocity; φ is the melt fraction; C is the composition, defined using the phase diagram described in the next section; I is the trace-element concentration; $\Delta \rho$ is the density contrast between melt and crystals and μ is the melt shear viscosity, both discussed below. Subscripts 's' and 'm' denote solid and melt, respectively. Primes denote the dimensionless equivalents of the dimensional variables.

The characteristic timescales and length scales used to non-dimensionalize the equations are given by¹⁰

$$\tau = \frac{1}{\Delta \rho_r g} \left(\frac{\mu_r \eta_0}{a^2 b} \right)^{\frac{1}{2}} \quad (6)$$

$$\delta = \left(\frac{\eta_0 a^2 b}{\mu_r} \right)^{\frac{1}{2}} \quad (7)$$

where $\Delta \rho_r$ is a reference density contrast and μ_r a reference melt shear viscosity discussed in below, g is the acceleration due to gravity, and the matrix viscosity is related to the melt fraction by^{10,34,50}

$$\frac{4}{3} \eta + \xi = \eta_0 \varphi^{-\beta} \quad (8)$$

where η is the shear viscosity, ξ is the bulk viscosity, η_0 is a reference shear viscosity and β is an adjustable constant. The permeability of the mush k_φ is given by^{10,34,48,51}

$$k_\varphi = a^2 b \varphi^\alpha \quad (9)$$

where a is the grain size, and b and α are adjustable constants.

Temperature and enthalpy are scaled using¹⁰

$$T' = \frac{T - T_s}{T_L - T_s} \quad (10)$$

$$h' = \frac{h - h_s}{h_L - h_s} \quad (11)$$

where the subscripts 'L' and 'S' denote liquidus and solidus respectively. The dimensionless scaling factor κ in equation (1) is given by¹⁰

$$\kappa = \frac{k_T \tau (T_L - T_S)}{\rho_r \delta^2 (c_p (T_L - T_S) + L_f)} \quad (12)$$

and the Stefan Number by¹⁰

$$\text{Ste} = \frac{L_f}{c_p (T_L - T_S) + L_f} \quad (13)$$

where k_T is the thermal diffusivity, c_p the sensible heat capacity, L_f the latent heat of fusion and ρ_r is a reference density discussed below.

The initial condition is chemically homogeneous crust with a constant linear geotherm T_{geo} and no melt present. Temperature, melt fraction and velocity are zero at the upper boundary (Earth surface); the lower boundary has constant T_{geo} and is set sufficiently deep that melt fraction and velocity remain zero. Equations (1)–(5) were solved numerically using finite difference methods and the MUSHREACT code that we developed. Equation (1) was approximated using a forward-time-centred-space scheme; equations (2) and (3) using a modified Lax–Wendroff scheme; and equation (4) using a centred scheme. Node spacing and time steps were chosen based on the results of convergence tests. Solutions reported here used 20–40 nodes per individual sill intrusion with time steps that were always less than the well known CFL condition^{34,71}.

The numerical methods and code have been validated extensively against analytical solutions^{10,34,51,71}. The energy conservation equation (1) is a special case of the general parabolic diffusion–advection equation, while the mass conservation equations (2) and (3) are special cases of the general hyperbolic flux conservative equation. Analytical solutions are available for simplified forms of these general equations, and the numerical methods were tested against these. An analytical solution is available for a simplified set of the compaction equations and the code was also tested against this.

Phase behaviour and composition-dependent material properties. Phase change and compositions are described using a binary eutectic phase diagram that approximates the behaviour of natural systems. Several previous studies of crustal igneous systems have used a similar approach, which is preferable to more complex models involving, for example, the thermodynamic software MELTS⁷² for two reasons. First, reactive flow leads to local changes in bulk composition, so the local phase equilibria must be recalculated at each location and time; this is trivial using a simple phase diagram, but computationally intensive (albeit possible) using MELTS. Second, it allows fundamental aspects of compositional evolution to be identified without the additional complexity associated with modelling the phase behaviour of natural systems^{10,37}.

Melt fraction is related to composition through

$$\varphi = \frac{C - C_s}{C_m - C_s} \quad (14)$$

where C is the local bulk composition. Assuming a linear release of enthalpy during melting, enthalpy is related to temperature through

$$h = c_p T + L_f \varphi \quad (15)$$

Using equations (14) and (15), and the temperature-dependent liquid and solid compositions determined from the phase diagram, the melt fraction is determined locally.

The binary eutectic phase diagram is described by a quadratic function given by¹⁰

$$C_m = \begin{cases} \frac{-a_2 - \sqrt{a_2^2 - 4a_1(a_3 - T)}}{2a_1} & C > e \\ \frac{-b_2 + \sqrt{b_2^2 - 4b_1(b_3 - T)}}{2b_1} & C < e \end{cases} \quad (16)$$

$$C_s = \begin{cases} 0 & C > e \\ 1 & C < e \end{cases} \quad (17)$$

Here, only compositions with $C < e$ were used. Values of the constants a_1 , a_2 and a_3 were selected so that the melting relations obtained for starting compositions chosen to represent crust and intruded basalt match typical experimental data for the equilibrium melting/crystallization of metagreywackes and basalt, respectively, over the pressure range 400–900 MPa (Extended Data Fig. 1a; Extended Data Table 1)^{12,73,74}. The match is surprisingly good given the simple phase behaviour adopted.

It is important to recognize that the static melt fraction versus temperature relations shown in Extended Data Fig. 1a are specific to the chosen starting bulk compositions. They are not valid if the bulk composition changes in response to reactive melt flow. The phase diagram provides a family of melting curves for all bulk compositions encountered in the reservoir; we show just two in Extended Data Fig. 1a. The effect of reactive melt flow in the reservoir is to decouple melt fraction and temperature (Extended Data Fig. 1c). High melt fraction can be found at low temperatures where reactive melt flow has caused the bulk composition of the mush to evolve towards the eutectic and vice versa. It is often assumed that high melt fraction necessitates high temperature^{6–8,22,23,31–33,75}. Reactive melt flow means that this is not the case in crustal mush reservoirs.

We choose to relate composition C to a simple measure of differentiation, the SiO_2 content, by

$$S_{\text{SiO}_2} = a_5 + a_6 \tanh(a_7 + a_8 C) \quad (18)$$

Values of the constants a_5 to a_8 were selected to yield a variation in SiO_2 content with temperature that matches melt SiO_2 content from the same experimental melting/crystallization data (Extended Data Fig. 1b; Extended Data Table 1). Again, the match is surprisingly good given the simple phase behaviour adopted.

Rearrangement of equations (14), (16) and (17), followed by substitution into equation (15), yields a cubic polynomial in melt fraction, dependent on enthalpy h and bulk composition C , which can be solved analytically¹⁰

$$\varphi = \frac{h}{L_f} - \frac{c_p}{L_f} \left[\frac{a_3 + \left(2a_1 \left(\frac{C - C_s + C_s \varphi}{\varphi} \right) + a_2 \right)^2 + a_2^2}{4a_1} \right] \quad (19)$$

The model includes partitioning of a trace element between crystals and melt. The concentration in the melt is given by

$$I_m = \frac{I}{K + \varphi(1 - K)} \quad (20)$$

and in the solid by

$$I_s = KI_m \quad (21)$$

In the cases modelled here, the intruding magma and crust have the same initial concentration of an incompatible trace element. This is unlikely to occur in nature, but allows the evolution of trace-element concentration in response to reactive melt flow in the mush to be more clearly observed and understood. Trace-element concentration does not affect the evolution of temperature or melt fraction, so the other key model results remain unchanged.

The density of the melt and matrix, and the viscosity of the melt, both vary as a function of composition. Solid and melt densities are given by

$$\rho_m = C \rho_{m,\text{min}} + (1 - C) \rho_{m,\text{max}} \quad (22a)$$

$$\rho_s = C \rho_{s,\text{min}} + (1 - C) \rho_{s,\text{max}} \quad (22b)$$

where the subscripts 'max' and 'min' denote, respectively, the most evolved and least evolved (most refractory) compositions in the system. The average density of the crystals plus melt mixture (mush or magma) is given by

$$\bar{\rho} = \phi \rho_m + (1 - \phi) \rho_s \quad (23)$$

The dimensionless density is obtained by dividing by a reference density ρ_r chosen as the initial density of the crust, and the dimensionless density contrast is obtained by dividing by a reference density contrast $\Delta \rho_r$ chosen to be the difference in density between the most refractory crystals ($\rho_{s,\text{max}}$) and most evolved melt ($\rho_{m,\text{min}}$).

The logarithm of melt shear viscosity μ is linearly related to the dimensionless silica content of the melt s_{SiO_2}

$$\mu = 10^{(\mu_{\text{max}} - \mu_{\text{min}}) s_{\text{SiO}_2} + \mu_{\text{min}}} \quad (24)$$

with

$$s_{\text{SiO}_2} = \frac{S_{\text{SiO}_2} - S_{\text{SiO}_2}^{\text{min}}}{S_{\text{SiO}_2}^{\text{max}} - S_{\text{SiO}_2}^{\text{min}}} \quad (25)$$

where S_{SiO_2} is the silica content of the melt (in weight per cent)⁵⁸. The dimensionless melt shear viscosity is then obtained by dividing by a reference viscosity μ_r

chosen to be the maximum melt viscosity in the system (corresponding to the most evolved composition), to yield

$$\mu' = 10^{(\log(\mu_{\min}/\mu_{\max})(1-s_{\text{SiO}_2}))} \quad (26)$$

In the illustrative models shown here, melt viscosity varies from a minimum of 1 Pa s to a maximum of 10^5 Pa s (Extended Data Table 1) for the most mafic and most felsic compositions respectively, which is reasonable for melt containing a few weight per cent water⁵⁸. A range of maximum melt viscosities is investigated in the sensitivity analysis below.

Modelling of sill intrusion. The governing equations do not include terms representing addition of heat and mass in response to repetitive sill intrusion. Each sill intrusion is modelled numerically, using a simple approach in which new nodes, populated with the properties (enthalpy, melt fraction, major element composition and trace-element concentration) of the magma in the sill, are added into the model at the target intrusion depth^{22,32,34}. The number of new nodes is chosen to yield the desired sill thickness. Pre-existing nodes below the location of sill intrusion are shifted downwards to accommodate the new nodes representing the sill; this approach represents, numerically, the way that intrusion of each new sill causes downward displacement of deeper crust and approximates isostatic equilibrium. Intrusion of each sill is assumed to occur over a timescale that is small compared to the thermal and chemical evolution of the magma reservoir and within a single time-step in the model. We note that injection of magma may generate local over-pressure, fracturing and, during the growing and active phases of the magma reservoir, locally disrupt the mush. Future refinements will focus on methods to better couple thermal and mechanical models.

Sill intrusion depth. Previous numerical studies have modelled repetitive intrusion by over-accretion, in which each new sill is intruded immediately above the previous sill; under-accretion, in which each new sill is intruded immediately below the previous sill; and random intrusion of sills and dykes around a fixed depth^{21–23,32–34}. The approach used here to link sill intrusion depth to the state of the mush reservoir at the time of intrusion yields variations in intrusion depth that are not captured by these previous models.

Controls on the depth of sill intrusions include rigidity contrasts and rheology anisotropy, resulting from variations in lithology and (if present) melt fraction; rotation of deviatoric stress such that the minimum deviatoric stress becomes vertical; and density contrasts between the surrounding country rock and intruding magma^{67,68}. The initial intrusion depth is chosen here to match the depth of an observed magma reservoir. Understanding why a sill should be initially emplaced at a given depth is beyond the scope of this Letter. Once the first sill is emplaced, the depth of subsequent intrusions is controlled by the density contrast between the intruding magma and the surrounding reservoir. The next sill intrusion occurs at the deepest level of the mush that has a lower bulk density (crystals plus melt) than the intruding magma. The top of the resulting reservoir tends to be close to the initial intrusion depth.

Density contrasts are controlled by the local composition and/or melt fraction of the mush reservoir. We use density contrasts here as a proxy for rigidity contrasts resulting from changes in rock composition or mush melt fraction^{21,56,57,67,68}. Density is calculated using equations (22) and (23); the chosen values of density for refractory crystals ($\rho_{s,\text{max}}$) and most evolved melt ($\rho_{m,\text{min}}$) (Extended Data Table 1) yield densities of about $3,000 \text{ kg m}^{-3}$ and about $2,600 \text{ kg m}^{-3}$ for solid basalt and evolved (felsic) rock compositions respectively, and densities of about $2,800 \text{ kg m}^{-3}$ and about $2,350 \text{ kg m}^{-3}$ for their corresponding molten counterparts. These values are consistent with measured data^{25,76,77}. The initial (reference) density of the solid crust is about $2,850 \text{ kg m}^{-3}$, consistent with data for intermediate rocks⁷⁷.

During the incubation phase, melt fraction falls to zero between successive sill intrusions (Extended Data Fig. 2), but variations in density arise in response to differentiation within each intruded sill as it cools. Differentiation yields a lower-density, evolved top and a higher-density, more refractory base (Extended Data Fig. 3a and Supplementary Video 1). Similar compositional trends are observed in sills now exposed at the surface^{10,78,79}. The density-controlled intrusion depth of each new sill is, therefore, located below the deepest evolved top of a previous, now solidified, sill intrusion.

During the growing and active phases of the reservoir (Extended Data Fig. 2b), melt is persistently present and the compositional and density variations formed during the incubation phase are reduced by reactive melt flow (Fig. 1 and Supplementary Video 1). Variations in density are then primarily controlled by melt fraction, so that the density-controlled intrusion depth of each new sill is located below the deepest high melt fraction layer.

Field observations from deep crustal sections suggest that intrusions progressively accumulate to form a mush zone^{16,21,76}. At early times, when the heat content of the reservoir is still low, intrusions cool without causing extensive melting of the surrounding crust, leaving septa of crust interleaved with the intruded sills. We model this by intruding sills at random over a range of 300 m above and below the intrusion depth determined by density contrast.

Random intrusion preserves septa of crustal rock between sill intrusions, whereas strictly density-controlled intrusion does not (see also Solano et al.¹⁰). Varying the depth range of random intrusion affects the frequency and volume of preserved septa, but does not otherwise substantially affect the results obtained. Although septa between early intrusions are preserved, septa between later intrusions, when the heat content of the reservoir is higher, are partially assimilated into the melt phase, causing crustal contamination of the melt^{16,76,80,81}.

Validity of the model at high melt fraction. The reactive flow and compaction formulation is applied in the model regardless of local melt fraction. However, it is strictly valid only when the crystals form a solid framework that will expel melt if it undergoes mechanical disruption or viscous deformation⁸². Estimates of the melt fraction at which this framework forms vary widely (over the range 0.4–0.7) and probably depend on local shear stresses and strain rates, and the crystal morphology and size distribution^{40,45,82–84}. Melt fractions higher than this are present in each sill immediately after intrusion and in the melt layers that form in response to reactive flow. However, we argue below that the formulation captures enough of the physics to yield informative results.

High melt fractions are present in the intruding sills over very short timescales (of the order of hundreds of years) because the sills cool very rapidly, losing heat to the surrounding reservoir and/or crust (for example, Extended Data Fig. 2a). Over these short timescales following each intrusion, crystal–melt separation is assumed in the model to occur only by reactive flow and compaction, omitting other mechanisms of crystal–melt separation⁸²; moreover, it is assumed that there is no bulk flow of melt and crystals driven by convection^{28,85,86}. However, the modelled cooling timescale is correct, because the rate of heat loss from each sill is dominated by conduction and this is described by equation (1)^{71,86}. Furthermore, in each sill, the model captures enough crystal–melt separation to yield a more evolved top, relatively enriched in incompatible trace elements, and a more refractory base, relatively depleted in incompatible trace elements, consistent with observations (for example, Fig. 3)^{10,78,79}.

High melt fractions are also present in the layers that form in response to reactive melt flow (for example, Fig. 1). These layers are persistently present once formed and the model again assumes that crystal–melt separation in each layer occurs only by reactive flow and compaction and that there is no bulk flow of melt and crystals driven by convection. However, the rate of delivery of new melt into the layer is controlled by reactive flow and compaction of the underlying mush where the formulation is valid. Moreover, the rate of upward movement of the layer, which affects cold storage, is controlled by the rate of upward movement of the solidus isotherm; this depends on conductive heat transfer in the overlying mush and solid rock, and is captured by the formulation. Thus we argue that the model captures the overall growth and upward migration of the layers.

Within each high-melt-fraction layer, the formulation probably does not correctly capture the variation in melt fraction. However, the modelled temperature in each layer is constant at the solidus; melt fraction is also high and approximately constant, controlled primarily by the local bulk composition (for example, Supplementary Video 1; Fig. 1). Thus, the modelled temperature and melt fraction assuming reactive flow with no bulk flow of melt and crystals are similar to what would be observed for the opposite end-member model of vigorous convection in which crystals are suspended and mixed in the magma²⁸. We argue that vigorous convection may be more likely given the results of earlier studies of single sill intrusions^{28,85,86}.

Magma mobilization. Accumulation of melt creates a high-melt-fraction layer which, as it migrates upwards, can remobilize old mush by causing a rapid increase in melt fraction. The short timescale of this process may not allow for local chemical equilibrium to be maintained, so older crystals can be preserved in the younger magma. Disequilibrium between melt and crystals may also give rise to resorption and zonation of crystals which is not described here.

The model does not attempt to capture migration out of the reservoir of the high-melt-fraction (low-crystallinity) magmas in the layer. Felsic magma that accumulates at the top of the reservoir is buoyant relative to the surrounding mush reservoir, so there is a pressure gradient to drive ascent to higher crustal levels or eruption at the surface^{25,26}. The magma in each sill also evolves during cooling to become more buoyant relative to the more refractory mush, which may drive ascent of less evolved magmas. Preliminary work, not reported here, suggests that removal of felsic magma accumulating in a high-melt-fraction layer at the top of the reservoir does not affect the formation of subsequent layers, so long as ongoing sill intrusions continue to supply new magma to the reservoir.

This preliminary work is not reported because the model does not yet include clearly defined criteria for magma removal and ascent. Moreover, we note that the presence of volatile species, such as H_2O , whose solubility is pressure-dependent, complicates phase relations and physical properties during magma ascent, and consequently is not considered here. Further work should determine the controls on mobilization and eruption of the low-crystallinity magmas present in crustal mush reservoirs. What is clear from the results obtained here is that the

compositions of low-crystallinity magmas that can leave the reservoir, regardless of how or why that happens, are bimodal. In our model, the melt composition evolves to the eutectic; in more chemically complex systems, melt composition will evolve to other low-variance states, such as cotectics or peritectics (reaction boundaries). In all cases, the effect is to buffer chemically the composition of accumulated melts, as recently suggested on the basis of phase-equilibrium experiments⁸⁷.

Magmatic systems at shallower depth. The results shown in Fig. 1 (and also in Supplementary Video 1 and Extended Data Fig. 3) illustrate the key processes occurring within a crustal mush reservoir and were obtained using values of the model parameters that are typical of crustal systems (Extended Data Table 1 and associated references). The initial intrusion depth was chosen to allow model results to be compared against a deep crustal section now exposed at the surface: the Upper Mafic Complex of the Ivrea-Verbano zone, Italy^{16,21,76}. The complex is interpreted to represent about 8 km of basalt intruded into the crust over a few million years (that is, at intrusion rates of the order of a few millimetres a year). The top of the complex is interpreted to have been located at a depth of about 18 km at the time of formation.

The model results can explain a wide range of magmatic phenomena. However, we recognize that many of the magmatic systems that provide compelling evidence for these phenomena cannot be approximated by a model tuned specifically to match data from the Upper Mafic Complex. In particular, systems providing evidence for cold storage and compositional bimodality are often located at shallower levels in the crust^{6–9}. Moreover, major- and trace-element and isotopic data for these systems suggest they may be supplied by magmas of more evolved composition than basalt^{6–9,88}. In transcrustal magmatic systems¹ there are probably multiple zones of intrusion: primitive basalt magmas may form intrusions deep in the crust that generate more evolved magmas; these magmas ascend through the crust to form intrusion zones at shallower depths.

We test here whether similar results are obtained if the first sill is intruded at a depth of 10 km rather than 18 km. Numerous magmatic systems are observed in geophysical data at similar depth^{2,4,5,38}. All model parameters are the same as used previously (Extended Data Table 1), except that we assume the initial geotherm is appropriate for thermally mature crust where, for example, a deeper magmatic zone has thermally primed the upper crust before the onset of shallower magmatism. Previous studies have shown that this is necessary to allow upper-crustal magmatic systems to form without a prohibitively long incubation period or unreasonably high rate of magma intrusion²³.

The results obtained are qualitatively similar to those observed at 18 km depth. There is an incubation period, during which the melt fraction rapidly falls to zero, with compositional contrasts formed by chemical differentiation within each sill intrusion before solidification, causing the intrusion depth to increase progressively (Supplementary Video 2; Extended Data Fig. 4a). During the growing phase, buoyant melt again migrates upwards through the mush and reactive melt flow reduces, or removes, early formed compositional contrasts, so that the intrusion depth becomes controlled by the locally varying melt fraction (Supplementary Video 2; Extended Data Fig. 5a).

During the active phase, the reservoir can again produce evolved, low-crystallinity magmas from the high-melt-fraction layer that forms beneath the solidus isotherm, close to the top of the reservoir (Supplementary Video 2; Extended Data Fig. 5b). When intrusion of new sills ends, the reservoir enters the waning phase (Supplementary Video 2; Extended Data Fig. 4b) until the mush has completely solidified.

Cold storage is again observed where upwardly migrating, evolved melt rapidly accumulates around older antecrysts derived from crystallization of early sills (Extended Data Fig. 6). In this shallower example, melt accumulation forms a low-crystallinity magma a few hundreds of years after the local temperature exceeds the solidus, but the magma contains antecrysts formed about 1.3 Myr earlier (Extended Data Fig. 6). Compositional bimodality is again observed, as magmas in the high-melt-fraction layers have evolved composition, but magmas in the sills shortly after intrusion have compositions close to that of the intruded basalt (Extended Data Fig. 7a). Thus, the key results are consistently observed in models of shallower magmatic systems created and sustained by basaltic magmatism.

Intrusion of more evolved magma. We now test whether similar results are obtained at 10 km if the intruding sills contain magma of intermediate (andesitic) rather than basaltic composition. All other model parameters are the same as used in the previous 10 km model (Extended Data Table 1). We do not model intrusion of rhyolite magma because our density-controlled intrusion depth model does not apply for rhyolite magma: density-controlled intrusion alone would suggest that rhyolite should mostly erupt. That evolved, low-density magmas often intrude rather than erupt has been a challenge to density-driven models of magma ascent and intrusion for many years^{67,68}.

Intrusion of intermediate-composition (about 61% SiO₂) magma yields qualitatively similar behaviour to that observed in response to intrusion of basaltic magma. The incubation, growing, active and waning phases of reservoir life are

all observed and, during the active phase, a high-melt-fraction layer containing evolved (felsic) magma overlies a thicker, low-melt-fraction mush (for example, Extended Data Fig. 8). Older antecrysts are again rapidly remobilized by the arriving melt layer although, in this case, storage is 'cool' rather than 'cold': the temperature remains above the solidus, but the melt fraction remains low until the melt layer arrives. Whether crystals are kept in cold (sub-solidus) or cool (supra-solidus) conditions may be difficult to determine from crystal chemistry data, requiring accurate estimates of reservoir and solidus temperatures^{6–9}; the key point is that the crystals are stored at low (non-eruptible) melt fraction, as opposed to 'warm storage' where the magma remains eruptible¹⁸.

Compositional bimodality is again observed, but here the magma compositions are either evolved (felsic), reflecting melt accumulation in the upwardly migrating layer, or intermediate, reflecting the composition of the intruding magma (Extended Data Fig. 7b). In general, we suggest that crustal mush reservoirs deliver magmas with compositions that reflect either (1) low-variance states, such as eutectics, cotectics or peritectics (reaction boundaries)⁸⁷ or (2) the intruding magma that creates the reservoir.

Intrusion depth model. Numerical tests show that compositionally bimodal, low-crystallinity magmas are obtained regardless of whether the intrusion depth is modelled using our sill intrusion depth model or simple under- or over-accretion. 'Cold' (or at least 'cool') storage of crystals, in a non-eruptible state, is also observed (for example, Extended Data Fig. 9a, b), except when intrusion depth is modelled using simple over-accretion. Over-accretion cannot yield cold or cool storage of antecrysts formed as part of the same magmatic event, as persistent sill intrusion at the top of the magma reservoir causes the melt layer to migrate upwards and form in the overlying crust (for example, Extended Data Fig. 9c and d). The crystals here are rapidly mobilized by the arrival of the melt layer, but the history of the crystals and their genetic relationship with the magmatic event may be much more complex. However, simple over-accretion requires the magma supplying each sill to pass through the mush reservoir regardless of local melt fraction, rheology or density, which is inconsistent with available evidence and models^{67,68}. We argue that our sill intrusion model better captures the effect of the local mush state on intrusion depth.

Sensitivity analysis. Extended Data Table 1 shows that crustal magma reservoirs are described by a broad range of material properties. Values of many of these are poorly constrained. A simple sensitivity analysis was used to confirm that the results obtained are typical.

Previous work has shown that solutions to equations (1)–(5) are largely dictated by the value of the dimensionless scaling factor^{51,71} κ . The effect of varying the other dimensionless parameter Ste is much smaller. Other studies, confirmed by additional numerical experiments conducted here, have shown that, for a given depth of intrusion and initial geothermal gradient, the thermal impact of intruding sills is controlled by the intrusion rate, irrespective of the model used to choose the sill intrusion depth^{22,23,32,33}. The chosen intrusion rate of 5 mm yr^{−1} for the example results shown here corresponds to the time-averaged magma productivity in arc settings simplified to a one-dimensional geometry²⁴. We now explore a range of intrusion rates around this value, consistent with estimates for different crustal magmatic systems and previous studies^{21–24,32–34}.

A simple Monte Carlo analysis⁸⁹ shows that 90% of the calculated values of κ for typical crustal parameters lie within the range $0.028 < \kappa < 2,160$ ($-1.55 < \log \kappa < 3.335$; see Extended Data Fig. 10a). Numerical solutions were obtained for ten values of $\log \kappa$ sampled evenly over this range in log space, for a range of values of the sill intrusion rate, three different intrusion depths and basalt that is intruded to a maximum thickness of 20 km (Extended Data Table 1). The results are summarized in Extended Data Fig. 10b and c by plotting the incubation time (the time required to reach the end of the incubation phase and produce a persistent mush reservoir; see Extended Data Fig. 2b), the activation time (the time required to produce an active reservoir with a low-crystallinity felsic magma layer; see Extended Data Fig. 2b), the bulk composition of the mobile magmas (that is, magmas with melt fraction >0.7), and the 'cold storage time' of antecrysts at the top of the reservoir, as a function of sill intrusion rate for the different intrusion depths. The 'cold storage time' is the time elapsed between the last intrusion and the local melt fraction exceeding 0.7 at locations close to the top of the reservoir (see Fig. 2 and Extended Data Figs. 6, 8). The cold storage time reflects the likely range of crystal ages in magmas that have achieved melt fractions exceeding 0.7.

The incubation time scales with the reciprocal of the intrusion rate squared, q^{-2} (Extended Data Fig. 10b). The same scaling has been obtained in previous, purely thermal, models of repetitive sill intrusion using a variety of intrusion depth schemes, showing that the incubation time is relatively insensitive to the details of sill intrusion^{22,23,90}. Varying the value of κ over the range specified has a negligible impact on the incubation time, regardless of intrusion rate or depth, reflecting the relatively small range of uncertainty in thermal parameters such as thermal conductivity, specific heat capacity and latent heat (Extended Data Table 1).

The impact of varying κ on the accumulation time is larger, especially at lower intrusion rates when the accumulation time may be several million years longer than the incubation time (Extended Data Fig. 10b). Longer incubation times are observed for large values of κ that correspond to larger values of the melt shear viscosity and smaller values of matrix grain size, and for small values of κ that correspond to large values of the matrix bulk viscosity (Extended Data Table 1). Nevertheless, the accumulation time is finite so long as the incubation time is reached within the maximum intruded thickness of basalt; in other words, the formation of a high-melt-fraction layer is inevitable, so long as a persistent mush reservoir is present.

The composition of the high-melt-fraction (eruptible) magma in the reservoir is always bimodal, irrespective of the value of κ , the intrusion rate or intrusion depth (Extended Data Fig. 10c). The magma in the intruded sills has a composition close to that of the intruding basalt, while the magma in the layer that accumulates at the top of the reservoir has an evolved (approximately eutectic) composition, consistent with the results shown earlier for specific cases (Fig. 3; Extended Data Fig. 7a).

The impact of varying κ on the cold storage time is larger, as is the effect of intrusion rate (Extended Data Fig. 10c). Smaller cold storage times are observed for larger values of κ that correspond to larger values of the melt shear viscosity and smaller values of matrix grain size (Extended Data Table 1). Smaller cold storage times are also observed for higher intrusion rates, because evolution of the system as a whole occurs more rapidly. The cold storage time reflects the relative timing of sill intrusion relative to remobilization. Nevertheless, the cold storage time is always non-zero; in other words, some antecrysts are stored in a cold (or cool), non-eruptible state, before remobilization by reactive flow.

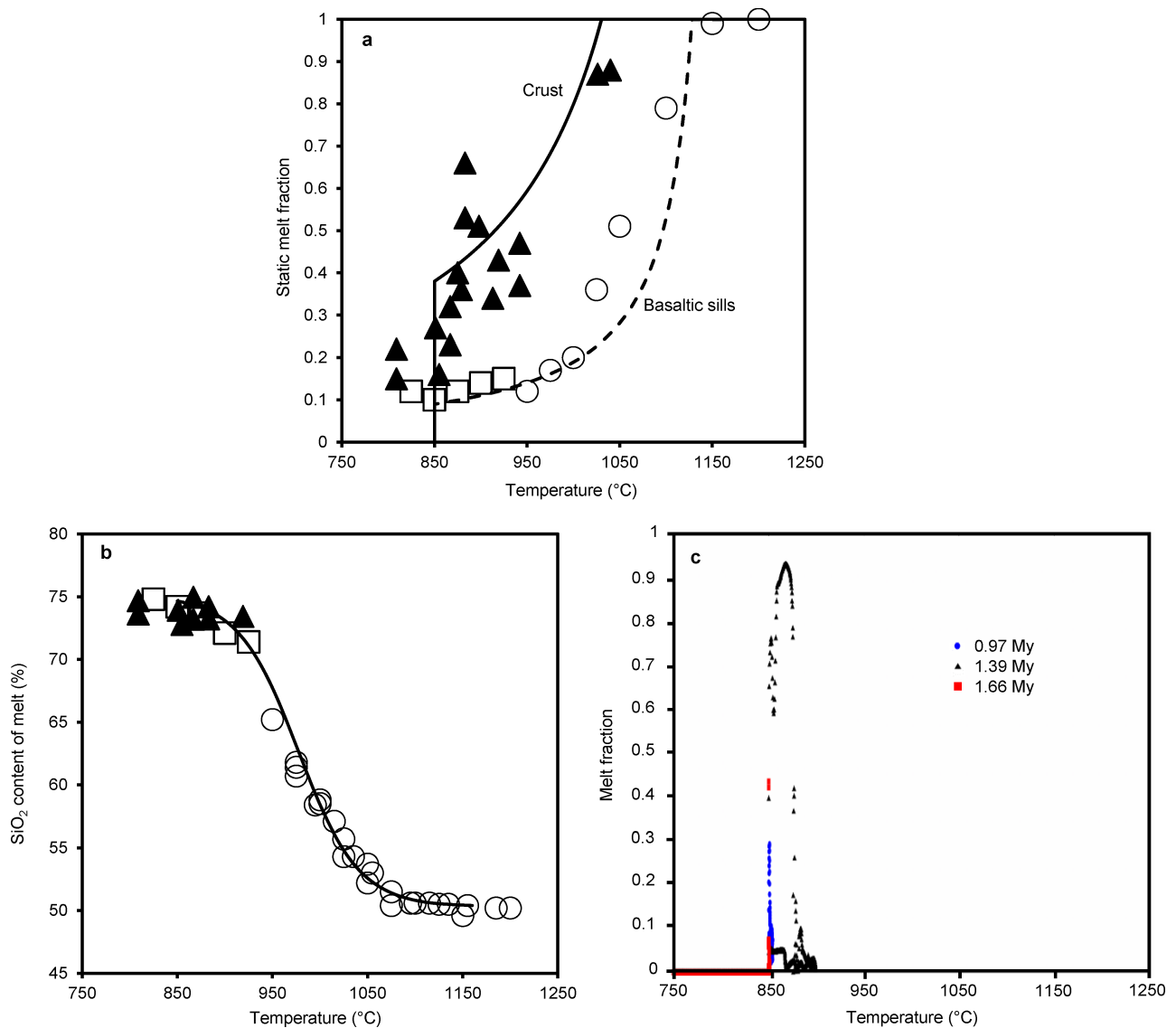
Code availability. The code (MUSHREACT) used to solve equations (1)–(5) and produce the results reported here is available from the corresponding author on request. The code is platform-dependent and is not optimized or tested for broad distribution, but the methodology is described within the article and preceding studies^{10,34}.

Data availability

No original data are reported that were not created using the software code (MUSHREACT). Data can be recreated using the code.

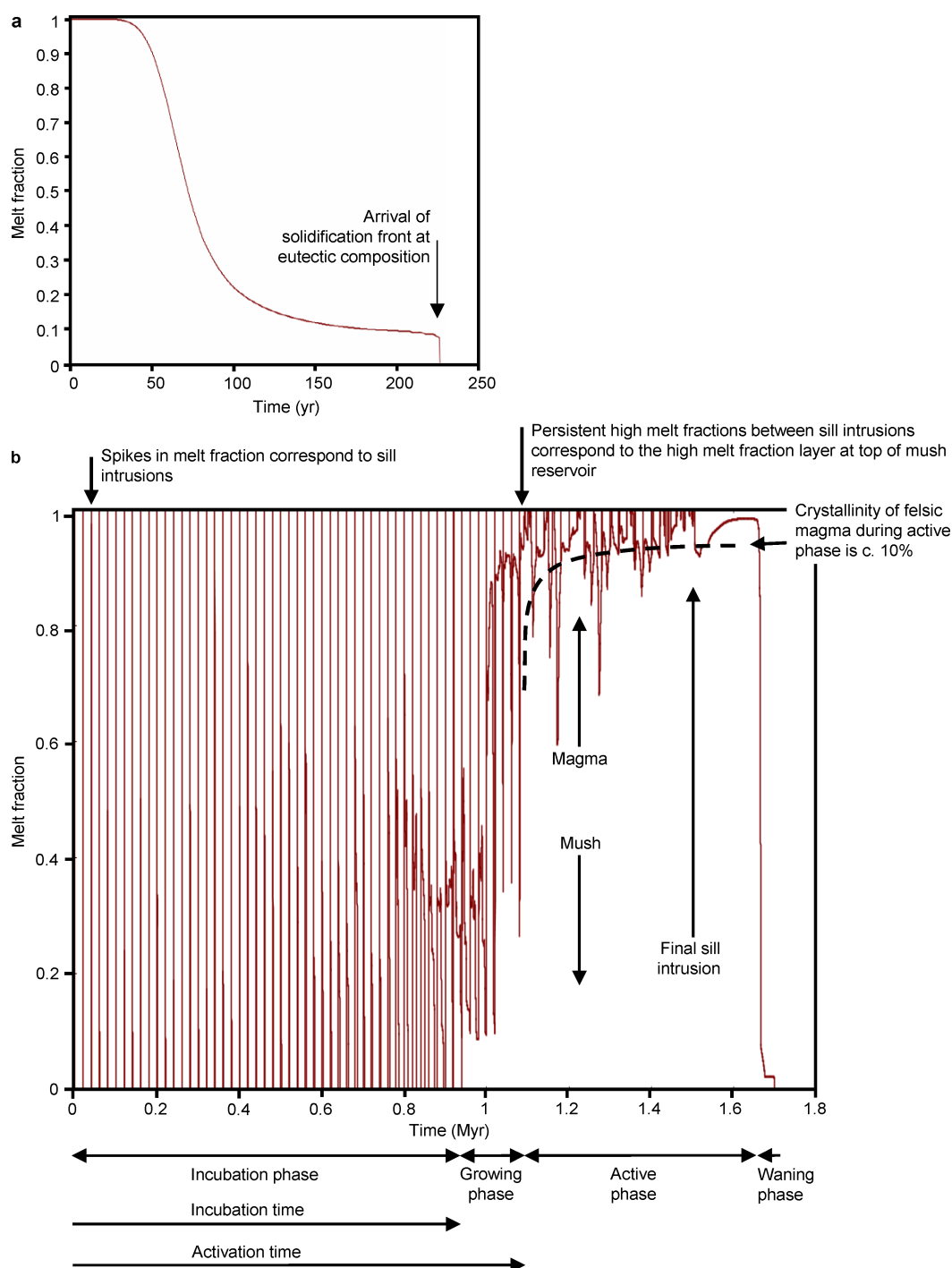
31. Hodge, D. S. Thermal model for origin of granitic batholiths. *Nature* **251**, 297–299 (1974).
32. Petford, N. & Gallagher, K. Partial melting of mafic (amphibolitic) lower crust by periodic influx of basaltic magma. *Earth Planet. Sci. Lett.* **193**, 483 (2001).
33. Dufek, J. & Bergantz, G. Lower crustal magma genesis and preservation: a stochastic framework for the evaluation of basalt–crust interaction. *J. Petrol.* **46**, 2167–2195 (2005).
34. Solano, J. M. S., Jackson, M. D., Sparks, R. S. J., Blundy, J. D. & Annen, C. Melt segregation in deep crustal hot zones: a mechanism for chemical differentiation, crustal assimilation and the formation of evolved magmas. *J. Petrol.* **53**, 1999–2026 (2012).
35. Bowen, N. L. *Evolution of the Igneous Rocks* 2nd edn, **362** (Dover, New York, 1956).
36. Hallworth, M. A., Huppert, H. E. & Woods, A. W. Crystallization and layering induced by heating a reactive porous medium. *Geophys. Res. Lett.* **31**, L13605 (2004).
37. Kerr, R. C., Woods, A. W., Grae Worster, M. & Huppert, H. E. Disequilibrium and macrosegregation during solidification of a binary melt. *Nature* **340**, 357–362 (1989).
38. Heise, W. et al. Melt distribution beneath a young continental rift: the Taupo Volcanic Zone, New Zealand. *Geophys. Res. Lett.* **34**, L14313 (2007).
39. Bachmann, O. & Bergantz, G. W. On the origin of crystal-poor rhyolites: extracted from batholithic crystal mushes. *J. Petrol.* **45**, 1565–1582 (2004).
40. Costa, A., Caricchi, L. & Bagdassarov, N. A model for the rheology of particle-bearing suspensions and partially molten rocks. *Geochem. Geophys. Geosyst.* **10**, Q03010 (2009).
41. Wolf, M. B. & Wyllie, P. J. Dehydration-melting of solid amphibolite at 10 kbar: Textural development, liquid interconnectivity and applications to the segregation of magmas. *Mineral. Petrol.* **44**, 151–179 (1991).
42. Ducea, M. N., Otamendi, J., Bergantz, G. W., Jianu, D. & Petrescu, L. in *Geodynamics of a Cordilleran Orogenic System: The Central Andes of Argentina and Northern Chile* (eds DeCelles, P. G. et al.) Geological Society of America Memoir Vol. 212, 125–138 (GSA, 2015).
43. Yoshino, T. & Okudaira, T. Crustal growth by magmatic accretion constrained by metamorphic P–T paths and thermal models of the Kohistan arc, NW Himalayas. *J. Petrol.* **45**, 2287–2302 (2004).
44. Hacker, B. R. et al. Reconstruction of the Talkeetna intraoceanic arc of Alaska through thermobarometry. *J. Geophys. Res.* **113**, B03204 (2008).
45. Rosenberg, C. L. & Handy, M. R. Experimental deformation of partially melted granite revisited: implications for the continental crust. *J. Metamorph. Geol.* **23**, 19 (2005).
46. Dell'Angelo, L. N., Tullis, J. & Yund, R. A. Transition from dislocation creep to melt-enhanced diffusion creep in fine-grained granitic aggregates. *Tectonophysics* **139**, 325–332 (1987).
47. Mei, S., Bai, W., Hiraga, T. & Kohlstedt, D. L. Influence of melt on the creep behavior of olivine-basalt aggregates under hydrous conditions. *Earth Planet. Sci. Lett.* **201**, 491–507 (2002).
48. McKenzie, D. P. The generation and compaction of partially molten rock. *J. Petrol.* **25**, 713–765 (1984).
49. Richter, F. M. & McKenzie, D. Dynamical models for melt segregation from a deformable matrix. *J. Geol.* **92**, 729–740 (1984).
50. Connolly, J. A. D. & Podladchikov, Y. Y. Compaction driven fluid flow in viscoelastic rock. *Geodin. Acta* **11**, 55–84 (1998).
51. Jackson, M. D., Cheadle, M. J. & Atherton, M. P. Quantitative modeling of granitic melt generation and segregation in the continental crust. *J. Geophys. Res.* **108**, 2332–2353 (2003).
52. Hersum, T. G., Marsh, B. D. & Simon, A. C. Contact partial melting of granitic country rock, melt segregation, and re-injection as dikes into ferrar dolerite sills, McMurdo dry valleys, Antarctica. *J. Petrol.* **48**, 2125 (2007).
53. Jackson, M. D., Gallagher, K., Petford, N. & Cheadle, M. J. Towards a coupled physical and chemical model for tonalite–trondhjemite–granodiorite magma formation. *Lithos* **79**, 43 (2005).
54. Keller, T., Katz, R. F. & Hirschmann, M. M. Volatiles beneath mid-ocean ridges: deep melting, channelised transport, focusing, and metasomatism. *Earth Planet. Sci. Lett.* **464**, 55–68 (2017).
55. Katz, R. F. Magma dynamics with the enthalpy method: benchmark solutions and magmatic focusing at mid-ocean ridges. *J. Petrol.* **49**, 2099–2121 (2008).
56. Ranalli, G. *Rheology of the Earth: Deformation and flow processes in Geophysics and Geodynamics* 2nd edn, 366 (Allen and Unwin, London, 1987).
57. Schmeling, H., Kruse, J. P. & Richard, G. Effective shear and bulk viscosity of partially molten rock based on elastic moduli theory of a fluid filled poroelastic medium. *Geophys. J. Int.* **190**, 1571–1578 (2012).
58. Giordano, D., Russell, J. K. & Dingwell, D. B. Viscosity of magmatic liquids: A model. *Earth Planet. Sci. Lett.* **271**, 123–134 (2008).
59. Bercovici, D., Ricard, Y. & Schubert, G. A two-phase model for compaction and damage. Part 1: general theory. *J. Geophys. Res.* **106**, 8887–8906 (2001).
60. Ricard, Y., Bercovici, D. & Schubert, G. A two-phase model for compaction and damage. Part 2: applications to compaction, deformation and the role of interfacial surface tension. *J. Geophys. Res.* **106**, 8907–8924 (2001).
61. Šrámek, O., Ricard, Y. & Bercovici, D. Simultaneous melting and compaction in deformable two-phase media. *Geophys. J. Int.* **168**, 964–982 (2007).
62. Simpson, G., Spiegelman, M. & Weinstein, M. I. A multiscale model of partial melts: 1. Effective equations. *J. Geophys. Res.* **115**, B04410 (2010).
63. Khazan, Y. Melt segregation and matrix compaction: the mush continuity equation, compaction/segregation time, implications. *Geophys. J. Int.* **183**, 601–610 (2010).
64. Karlstrom, L., Dufek, J. & Manga, M. Magma chamber stability in arc and continental crust. *J. Volcanol. Geotherm. Res.* **190**, 249–270 (2010).
65. Parmigiani, A., Faroughi, S., Huber, C., Bachmann, O. & Su, Y. Bubble accumulation and its role in the evolution of magma reservoirs in the upper crust. *Nature* **532**, 492–495 (2016).
66. Huppert, H. E. & Woods, A. W. The role of volatiles in magma chamber dynamics. *Nature* **420**, 493–495 (2002).
67. Menand, T. Physical controls and depth of emplacement of igneous bodies: a review. *Tectonophysics* **500**, 11–19 (2011).
68. Kavanagh, J. L., Boutelier, D. & Cruden, A. R. The mechanics of sill inception, propagation and growth: experimental evidence for rapid reduction in magmatic overpressure. *Earth Planet. Sci. Lett.* **421**, 117–128 (2015).
69. Spiegelman, M., Kelemen, P. & Aharonov, E. Causes and consequences of flow organization during melt transport: the reaction infiltration instability in compactible media. *J. Geophys. Res.* **106**, 2061–2077 (2001).
70. Liang, Y., Schiemenz, A., Hesse, M. A. & Parmentier, E. M. Waves, channels, and the preservation of chemical heterogeneities during melt migration in the mantle. *Geophys. Res. Lett.* **38**, L20308 (2011).
71. Jackson, M. D. & Cheadle, M. J. A continuum model for the transport of heat, mass and momentum in a deformable, multicomponent mush, undergoing solid-liquid phase change. *Int. J. Heat Mass Transfer* **41**, 1035–1048 (1998).
72. Ghiorso, M. S. & Sack, R. O. Chemical mass transfer in magmatic processes IV. A revised and internally consistent thermodynamic model for the interpolation and extrapolation of liquid–solid equilibria in magmatic systems at elevated temperatures and pressures. *Contrib. Mineral. Petrol.* **119**, 197–212 (1995).
73. Vielzeuf, D. & Montel, J. M. Partial melting of metagreywackes. Part I: fluid-absent experiments and phase relationships. *Contrib. Mineral. Petrol.* **117**, 375–393 (1994).
74. Blatter, D. L., Sisson, T. W. & Ben Hankins, W. Crystallization of oxidized, moderately hydrous arc basalt at mid- to lower-crustal pressures: implications for andesite genesis. *Contrib. Mineral. Petrol.* **166**, 861–886 (2013).
75. Burgisser, A. & Bergantz, G. W. A rapid mechanism to remobilize and homogenize highly crystalline magma bodies. *Nature* **471**, 212–215 (2011).
76. Sinigoi, S., Quick, J. E., Mayer, A. & Demarchi, G. Density-controlled assimilation of underplated crust, Ivrea-Verbano Zone, Italy. *Earth Planet. Sci. Lett.* **129**, 183–191 (1995).
77. Murase, T. & McBirney, A. R. Properties of some common igneous rocks and their melts at high temperatures. *Geol. Soc. Am. Bull.* **84**, 3563–3592 (1973).
78. Gibb, F. G. F. & Henderson, C. M. B. Convection and crystal settling in sills. *Contrib. Mineral. Petrol.* **109**, 538–545 (1992).
79. Latypov, R. M. The origin of basic–ultrabasic sills with S-, D-, and I-shaped compositional profiles by in-situ crystallization of a single input of phenocryst-poor parental magma. *J. Petrol.* **44**, 1619–1656 (2003).
80. Hildreth, W. & Moorbath, S. Crustal contributions to arc magmatism in the Andes of Central Chile. *Contrib. Mineral. Petrol.* **98**, 455 (1988).

81. Sisson, T. W., Salters, V. J. M. & Larson, P. B. Petrogenesis of Mount Rainier andesite: magma flux and geologic controls on the contrasting differentiation styles at stratovolcanoes of the southern Washington Cascades. *Geol. Soc. Am. Bull.* **126**, 122–144 (2014).
82. Holness, M. B. Melt segregation from silicic mushes: a critical appraisal of possible mechanisms and their microstructural record. *Contrib. Mineral. Petrol.* **173**, 48 (2018).
83. Philpotts, A. R. & Dickson, L. D. The formation of plagioclase chains during convective transfer in basaltic magma. *Nature* **406**, 59–61 (2000).
84. Castruccio, A., Rust, A. & Sparks, R. S. J. Rheology and flow of crystal-rich bearing lavas: insights from analogue gravity currents. *Earth Planet. Sci. Lett.* **297**, 471–480 (2010).
85. Worster, G. M., Huppert, H. E. & Sparks, R. S. J. Convection and crystallization in magma cooled from above. *Earth Planet. Sci. Lett.* **101**, 78–89 (1990).
86. Bergantz, G. W. & Dawes, R. in *Magmatic Systems* (ed. Ryan, M. P.) Ch. 13 (Academic, San Diego, 1994).
87. Blatter, D. L., Sisson, T. W. & Ben Hankins, W. Voluminous arc dacites as amphibole reaction-boundary liquids. *Contrib. Mineral. Petrol.* **172**, 27 (2017).
88. Tiell, C. B., Vazquez, J. A. & Boyce, J. W. Months between rejuvenation and volcanic eruption at Yellowstone caldera, Wyoming. *Geology* **43**, 695–698 (2015).
89. Sobol, I. M. *A Primer for the Monte Carlo Method* 1st edn, 126 (CRC Press, 1994).
90. Michaut, C. & Jaupart, C. Ultra-rapid formation of large volumes of evolved magma. *Earth Planet. Sci. Lett.* **250**, 38–52 (2006).



Extended Data Fig. 1 | Phase behaviour and compositions of the modelled system. **a**, Static melt fraction versus temperature for the modelled basalt and crust, extracted from the binary phase diagram for the chosen initial bulk compositions. Also shown are experimental equilibrium melting/crystallization data for metagreywackes and basalt over the pressure range^{12,73,74} 400–900 MPa. Triangles denote data from ref.⁷³; circles denote data from ref.⁷⁴; squares denote data from ref.¹². Static melt fraction denotes the melt fraction obtained if there is no relative motion of melt and crystals, so the bulk composition remains

constant. **b**, SiO₂ content versus temperature modelled here. Also shown are experimental data corresponding to those shown in **a**. **c**, Melt fraction versus temperature obtained from the numerical model (data extracted from Supplementary Video 1 at three snapshots in time (0.97 Myr, 1.39 Myr and 1.66 Myr after the onset of sill intrusions) corresponding to Fig. 1 and Extended Data Fig. 3b. Reactive flow in the mush decouples temperature and melt fraction: high melt fraction can be found at low temperatures where reactive melt flow has caused the bulk composition of the mush to evolve and vice versa.

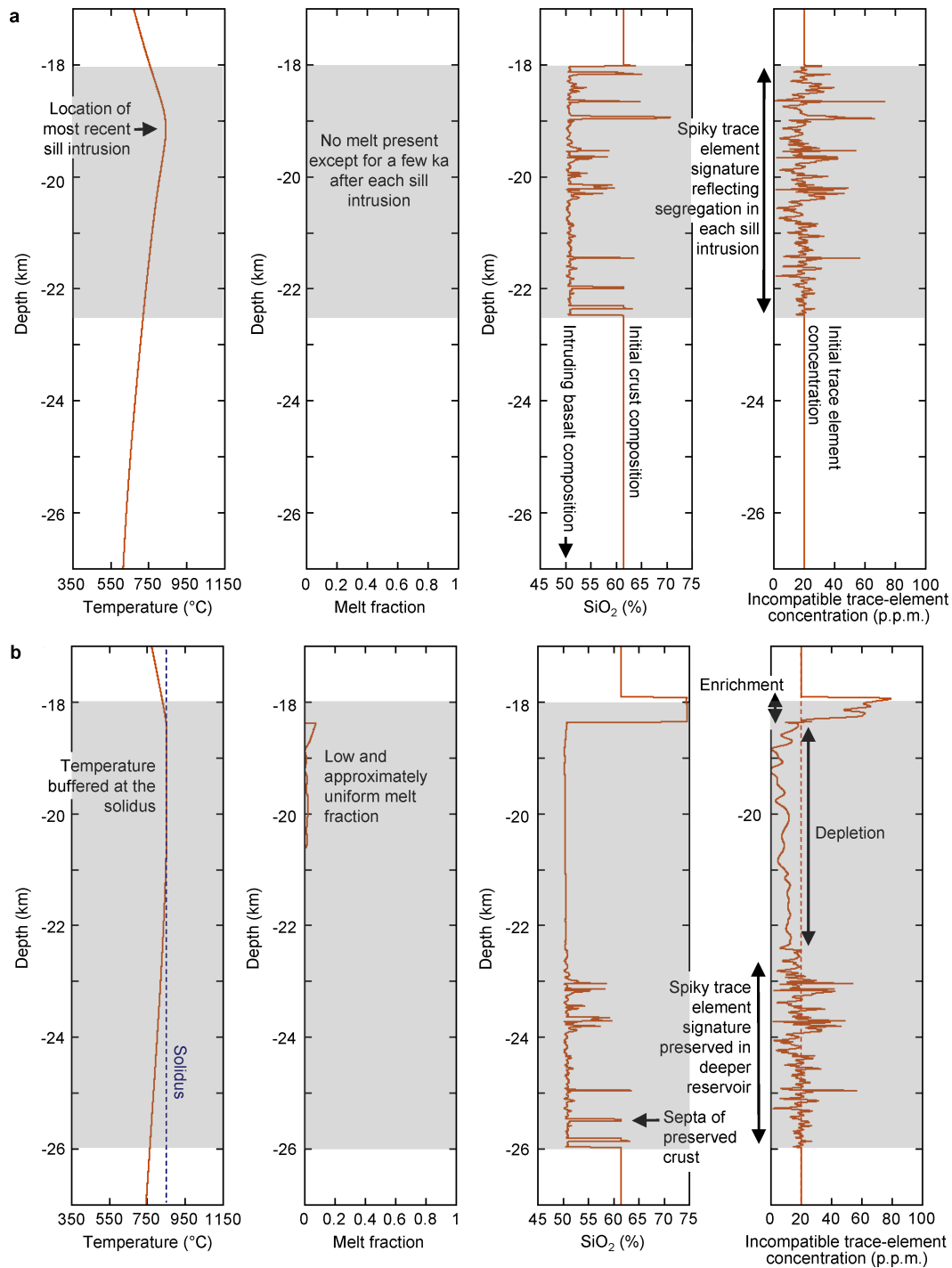


Extended Data Fig. 2 | Maximum melt fraction as a function of time.

a, Following a single sill intrusion during the incubation period.

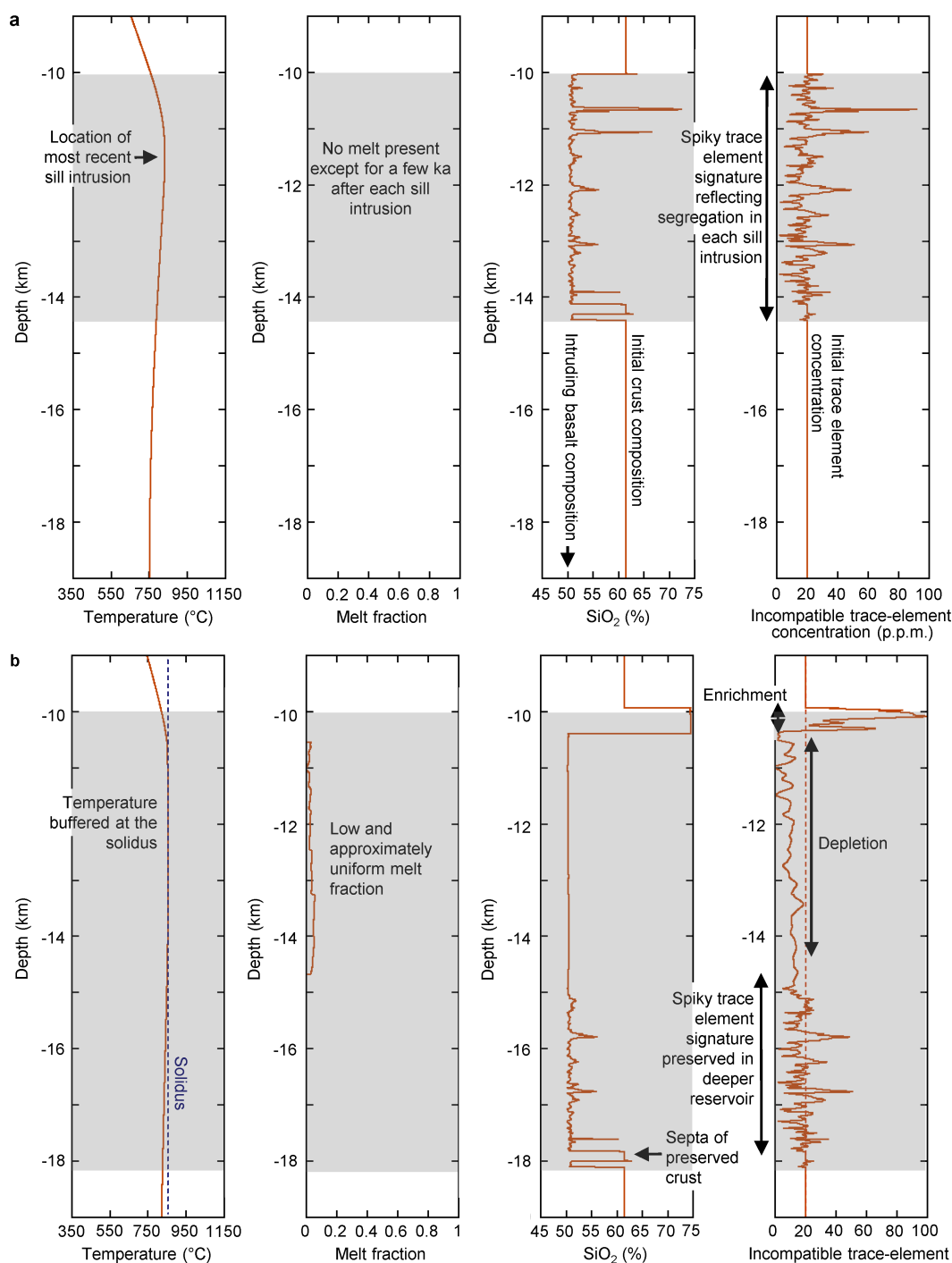
b, Over the life of the reservoir. In **a**, the sill cools rapidly, with the melt fraction falling below 0.7 (that is, the crystallinity exceeding 30%) within 63 years after intrusion, and the sill solidifying within 225 years. The sharp decrease in melt fraction before full solidification is physical and represents the arrival of the solidification front during crystallization at the eutectic. In **b**, during the 'incubation phase', maximum melt fraction spikes after each sill intrusion, but decreases rapidly and falls to zero between sill intrusions. The incubation phase ends when the melt fraction remains greater than zero between sill intrusions. During the 'growing

phase', the maximum melt fraction at the top of the mush reservoir increases in response to the reactive flow of buoyant melt. Spikes in melt fraction correspond to ongoing sill intrusions deeper in the reservoir. Melt fraction at the top of the mush increases until, during the 'active phase', evolved, low-crystallinity (<30%) magma is present, which is likely to leave rapidly and ascend to shallower crustal levels. New sill intrusions cease and, some time later, the melt fraction at the top of the mush also starts to decrease. Overall, the reservoir is cooling. This is the 'waning phase', at the end of which the reservoir has completely solidified. Data in both plots were extracted from Supplementary Video 1.



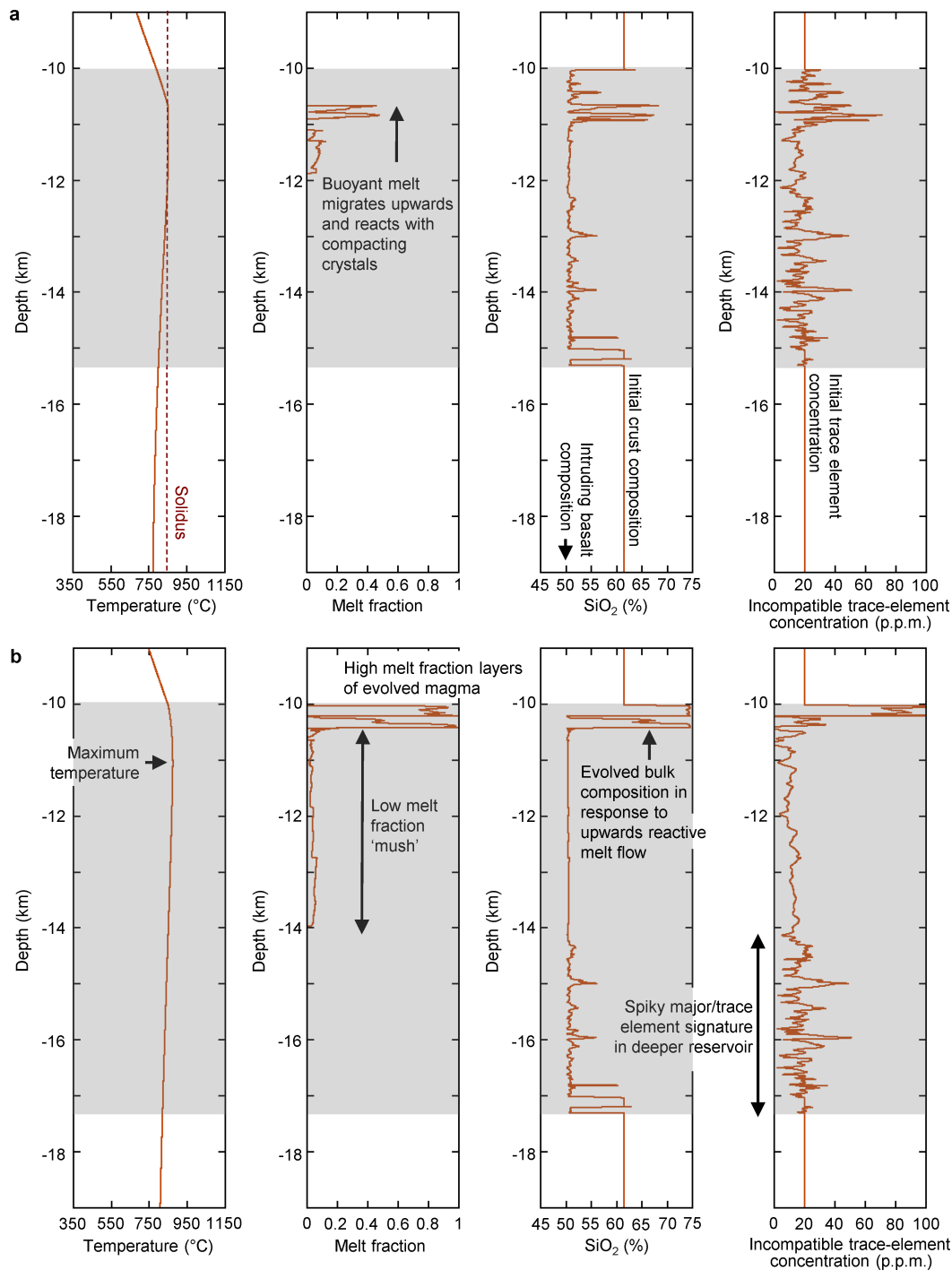
Extended Data Fig. 3 | Snapshots showing temperature, melt fraction, bulk composition and incompatible trace-element concentration as a function of depth through a crustal section at 18 km during the incubation and waning phases of the reservoir. a, 0.82 Myr after the onset of sill intrusions. b, 1.66 Myr after the onset of sill intrusions. Snapshots are taken from Supplementary Video 1. At early times, during the incubation phase (a), individual sills cool rapidly. During the growing phase (not shown here; see Fig. 1a), a persistent magma reservoir forms

but the melt fraction is low and relatively uniform. However, buoyant melt migrates upwards and begins to accumulate at the top of the reservoir. During the active phase (not shown here; see Fig. 1b), a high-melt-fraction layer forms. At late times, during the waning phase (b), sill intrusions cease and the mush cools and solidifies. The shaded areas in all plots denote the vertical extent of basalt intrusion at that time. Equivalent results for intrusions at 10 km depth are shown in Extended Data Fig. 4.



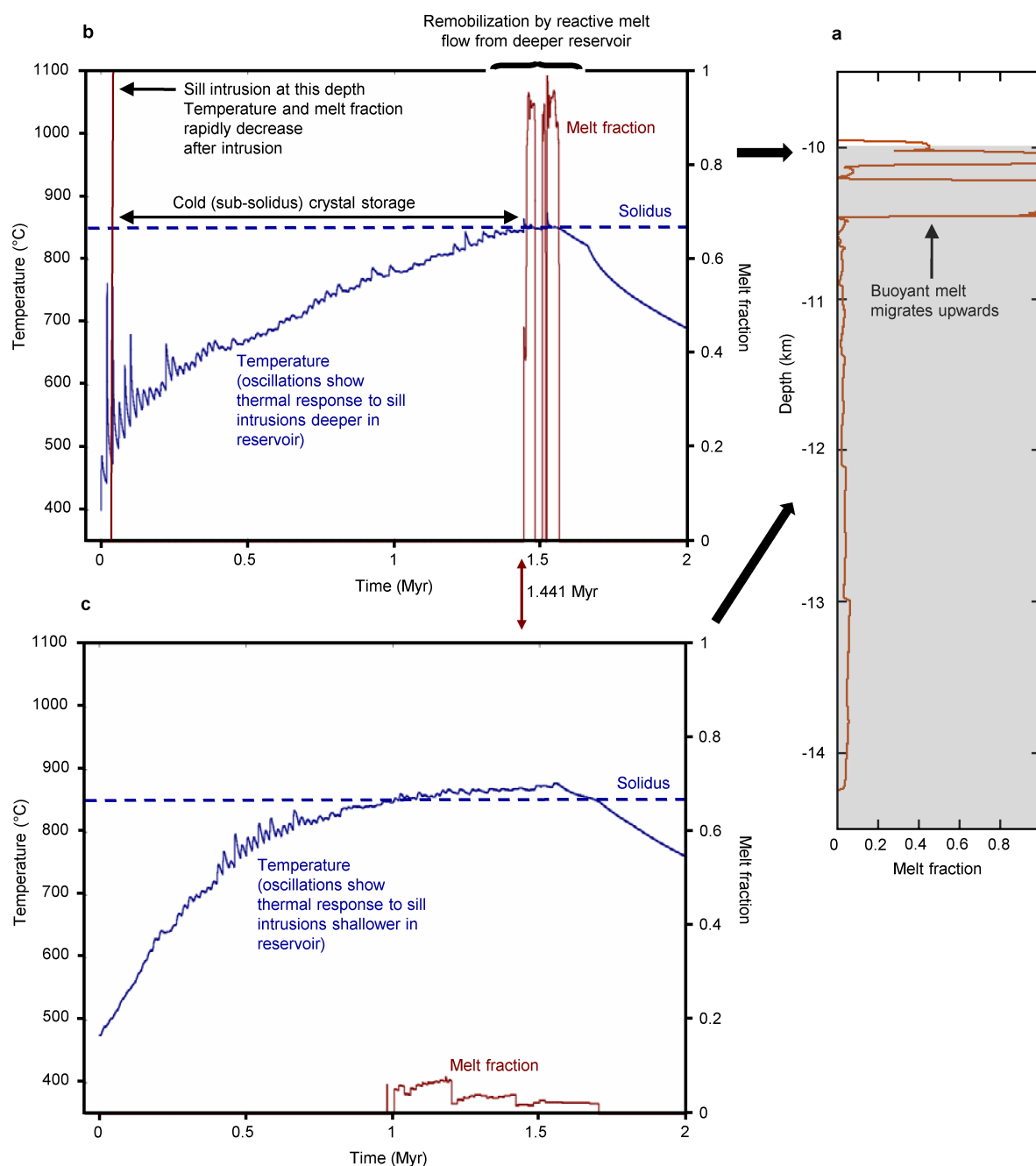
Extended Data Fig. 4 | Snapshots showing temperature, melt fraction, bulk composition and incompatible trace-element concentration as a function of depth through a crustal section at 10 km depth during the incubation and waning phases. a, 0.82 Myr after the onset of sill intrusions. b, 1.66 Myr after the onset of sill intrusions. Snapshots are taken from Supplementary Video 2. The results are qualitatively very

similar to those obtained at 18 km depth (Extended Data Fig. 3). During the incubation phase (a), individual sills cool rapidly. During the waning phase (b), sill intrusions cease and the mush cools and solidifies. The shaded areas in all plots denote the vertical extent of basalt intrusion at that time.



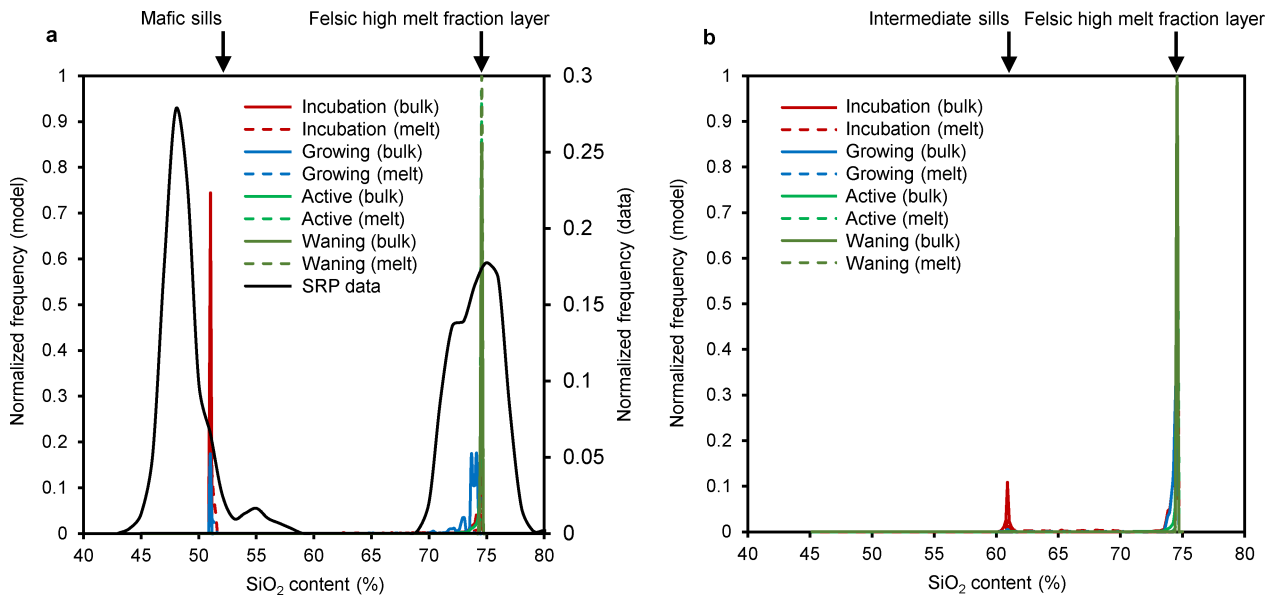
Extended Data Fig. 5 | Snapshots showing temperature, melt fraction, bulk composition and incompatible trace element concentration as a function of depth through a crustal section at 10 km depth during the growing and active phases. a, 0.99 Myr after the onset of sill intrusions. b, 1.39 Myr after the onset of sill intrusions. Snapshots are taken from Supplementary Video 2. The results are qualitatively very similar to those obtained at 18 km depth (Fig. 1). During the growing phase (a), a

persistent mush reservoir forms but the melt fraction is low. Buoyant melt migrates upwards and begins to accumulate at the top of the reservoir. During the active phase (b), the accumulating melt forms a high-melt-fraction layer containing mobile magma. The composition of the melt in the layer is evolved and enriched in incompatible trace elements. Elsewhere in the mush, the melt fraction remains low. The shaded areas in all plots denote the vertical extent of basalt intrusions at that time.



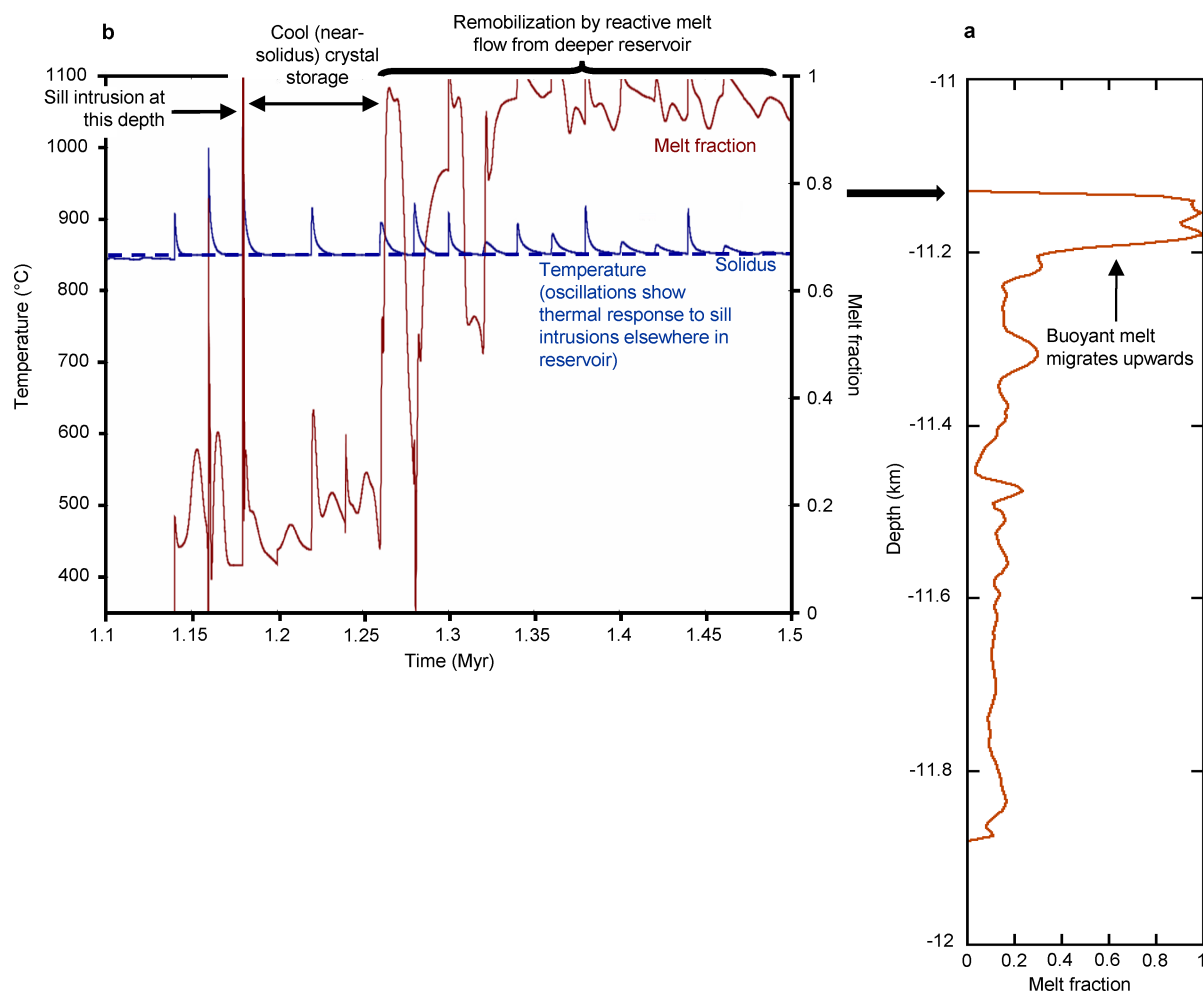
Extended Data Fig. 6 | Cold storage and rapid remobilization of magma in a reservoir at 10 km depth. Results are qualitatively very similar to those obtained at 18 km depth (Fig. 2). **a**, Melt fraction as a function of depth at the first snapshot after remobilization at 10 km (1.441 Myr after the onset of sill intrusions). The shaded area denotes intruded basalt. The reactive flow of buoyant melt produces a high-melt-fraction layer that migrates upwards. **b**, Temperature and melt fraction as a function of time at a depth of 10 km. Similar results are obtained over the depth range 10–10.5 km. Early sill intrusions rapidly cool and crystallize. The crystals are kept in ‘cold storage’ at sub-solidus temperature, but the temperature gradually increases in response to sill intrusions deeper in

the reservoir. Soon (<0.3 kyr) after the temperature exceeds the solidus, the high-melt-fraction layer arrives at this depth and the reservoir is remobilized: the melt fraction increases rapidly to form a low-crystallinity magma. The melt fraction increases much more rapidly and to a higher value than would be possible by melting alone. **c**, Temperature and melt fraction as a function of time at a depth of 12 km. Similar results are obtained over the depth range 10.5–15 km. Melt fraction remains low because reactive flow has left a refractory residue at this depth. There is no remobilization, despite the increase in temperature. Data were extracted from Supplementary Video 2.



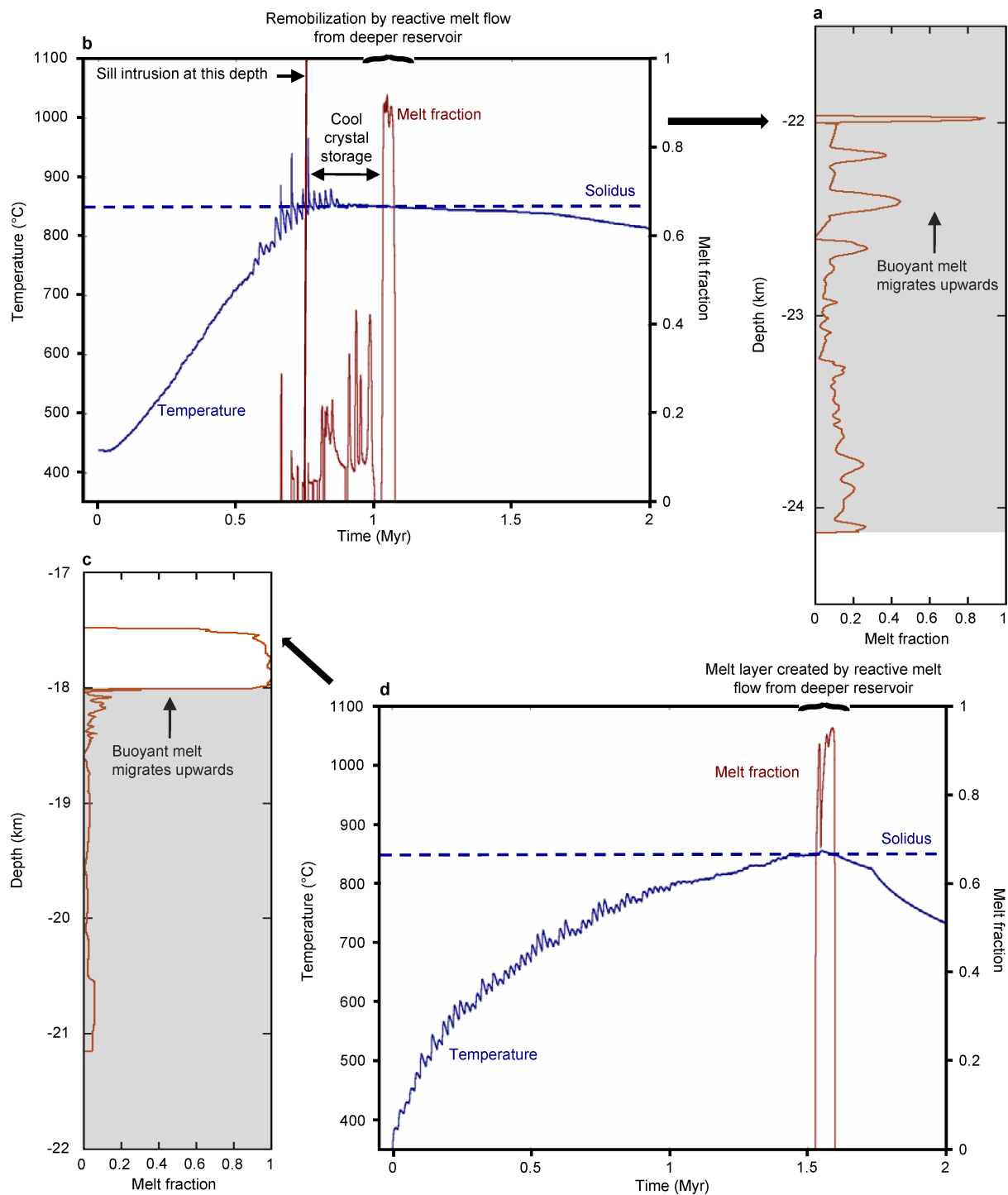
Extended Data Fig. 7 | Geochemical consequences of reactive melt flow in crustal magma reservoirs at 10 km depth. **a**, Intrusion of mafic sills; **b**, intrusion of intermediate sills. Both plots show SiO₂ content of low-crystallinity (crystal fraction <30%) magmas. Solid curves show bulk magma composition (melt plus crystals); dashed curves show melt composition alone. The peak at low SiO₂ corresponds to magma within the intruding sills; the peak at high SiO₂ corresponds to magma within

high-melt-fraction layers near the top of the reservoir. In **a**, measured data from the Snake River Plain (SRP) are shown for comparison²⁹; the bimodality is clear although the basalt has a lower SiO₂ content than modelled here. Bimodal compositions correspond to (1) the magma intruded into the reservoir, and (2) the most evolved composition obtained by differentiation.



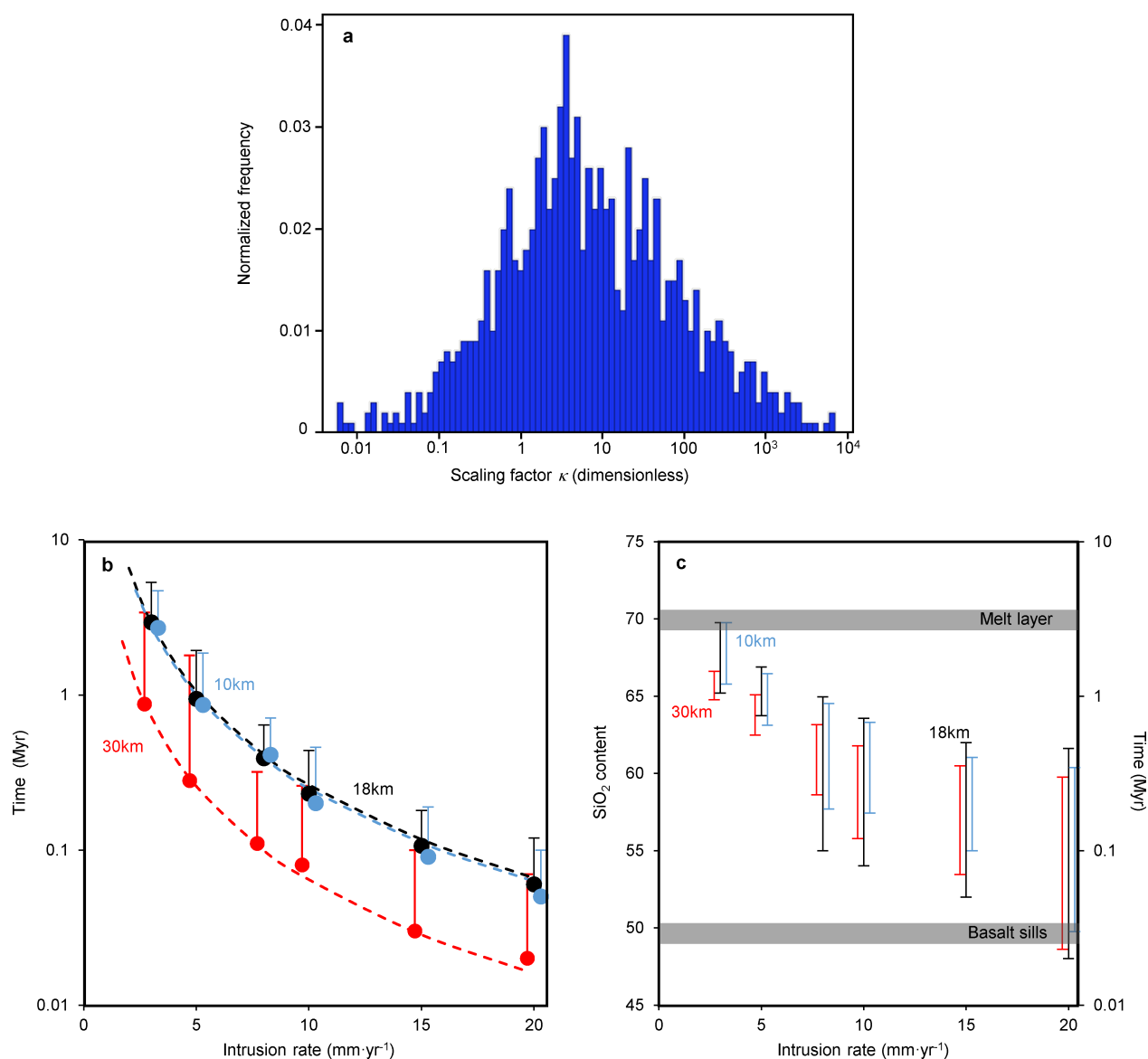
Extended Data Fig. 8 | Cool storage and rapid remobilization of magma in a reservoir created by intrusion of intermediate magma at 10 km depth. Results are qualitatively similar to those obtained by intruding basalt magma. **a**, Melt fraction as a function of depth at the first snapshot after remobilization at a depth of 11.4 km (1.28 Myr after the onset of sill intrusions). Reactive flow of evolved, buoyant melt produces a high-melt-fraction layer that migrates upwards. **b**, Temperature and melt fraction as a function of time at a depth of 11.4 km. Early sill intrusions rapidly

cool and crystallize. The crystals are kept in 'cool storage' at near-solidus temperature. At 1.28 Myr, the high-melt-fraction layer arrives at this depth and the reservoir is remobilized: the melt fraction increases rapidly to form a low-crystallinity magma. The melt fraction increases much more rapidly and to a higher value than would be possible by melting alone. Melt fraction deeper in the reservoir remains low because reactive flow has left a refractory residue at this depth.



Extended Data Fig. 9 | Consequences of emplacement during under- and over-accretion. **a**, Melt fraction as a function of depth during under-accretion, at the first snapshot after remobilization at a depth of 22 km (1.02 Myr after the onset of sill intrusions). Reactive flow of evolved, buoyant melt produces a high-melt-fraction layer that migrates upwards. **b**, Temperature and melt fraction as a function of time at a depth of 22 km during under-accretion. Similar results are obtained over the depth range 22–22.5 km. Under-accretion causes the sill intrusion depth to increase progressively from 18 km; in this case, an intrusion at 22 km occurs at 0.75 Myr ago that rapidly cools and crystallizes. The crystals are kept in ‘cool storage’ at a close-to-solidus temperature. At 1.02 Myr the

high-melt-fraction layer arrives at this depth and the reservoir is remobilized. **c**, Melt fraction as a function of depth during over-accretion, at a snapshot in time (1.53 Myr after the onset of sill intrusions). In this case, the high-melt-fraction layer has migrated into the overlying country rock. **d**, Temperature and melt fraction as a function of time at a depth of 17.5 km during over-accretion, close to the top of the active magma reservoir. Similar results are obtained over the depth range 17.5–18 km. Crystals in the magma are sourced from the country rock and may be genetically unrelated to the melt. There is no cold storage of crystals brought into the reservoir by basaltic sill intrusions, as intrusion occurs deeper in the reservoir. In **a** and **c**, the shaded area denotes intruded basalt.



Extended Data Fig. 10 | Sensitivity analysis. **a**, A frequency plot showing values of the dimensionless scaling factor κ calculated using equation (12). Values of the input values were varied uniformly over the range given in Extended Data Table 1 in a simple Monte Carlo analysis⁸⁹. **b**, Incubation and activation time; **c**, Cold storage time and eruptible magma composition. Error bars and shaded regions in **b** and **c** denote the

effect of varying the dimensionless scaling factor κ over the range $0.028 < \kappa < 2,160$. Error bars on the incubation time are within the symbol size. Dashed lines denote the fit to the incubation time of the form q^{-2} , where q is the intrusion rate. Colours in **b** and **c** denote different initial emplacement depths of 10 km, 18 km and 30 km. Models were run for a maximum 20 km of intruded basalt.

Extended Data Table 1 | Parameters used in the numerical experiments

| Symbol | Description and sources | Example case | Sensitivity analysis | Units |
|---------------------------|---|-----------------------------|---|---------------------------------------|
| k_T | thermal conductivity ^{22,23,33,51} | 3 | 1 - 3 | $W \cdot ^\circ C^{-1} \cdot m^{-1}$ |
| c_p | specific heat capacity ⁵¹ | 1100 | 1,020 - 1,220 | $J \cdot kg^{-1} \cdot ^\circ C^{-1}$ |
| L_f | latent heat ⁵¹ | 550000 | 400,000 - 600,000 | $J \cdot kg^{-1}$ |
| $T_L - T_S$ | liquidus-solidus interval ^{12,73,74} | 310 | 310 | $^\circ C$ |
| T_S | solidus ^{12,73,74} | 850 | 850 | $^\circ C$ |
| T_{geo} | initial geotherm ^{21-23,51} | 20 | 20, 40 | $^\circ C \cdot km^{-1}$ |
| a | matrix grain radius ⁵¹ | 2.75×10^{-3} | 5×10^{-4} - 5×10^{-3} | m |
| α | permeability exponent ⁵¹ | 3 | 3 | None |
| β | bulk viscosity exponent ⁵¹ | 0.5 | 0.5 | None |
| b | permeability constant ⁵¹ | 1/125 | 1/2500 - 1/50 | None |
| μ_{max} | shear viscosity of most evolved melt ⁵⁸ | 10^5 | 10^4 - 10^6 | $Pa \cdot s$ |
| μ_{min} | shear viscosity of least evolved melt ⁵⁸ | 1 | 1 | $Pa \cdot s$ |
| η_r | reference matrix shear viscosity ^{26,50,51,57} | 10^{15} | 10^{14} - 10^{17} | $Pa \cdot s$ |
| q | sill intrusion rate ^{21-24,32-34} | 5 | 1 - 20 | $mm \cdot yr^{-1}$ |
| z_s | sill thickness ^{21-23,42} | 100 | 50-200 | m |
| a_1, a_2, a_3 | phase behavior parameters | 50, -360, 1433.15 | 50, -360, 1433.15 | $^\circ C$ |
| a_4, a_5, a_6, a_7, a_8 | silica content modelling parameters | 62.7, 12.38, -0.0158, 15.44 | 62.7, 12.38, -0.0158, 15.44 | - |
| ρ | reference density ⁷⁷ | 2850 | 2850 | $kg \cdot m^{-3}$ |
| ρ_{smin} | density of most evolved solid composition ^{25,76,77} | 3000 | 3000 | $kg \cdot m^{-3}$ |
| ρ_{smax} | density of least evolved solid composition ^{25,76,77} | 2600 | 2600 | $kg \cdot m^{-3}$ |
| ρ_{mmin} | density of most evolved melt composition ^{26,76,77} | 2880 | 2880 | $kg \cdot m^{-3}$ |
| ρ_{mmax} | density of least evolved melt composition ^{25,76,77} | 2350 | 2350 | $kg \cdot m^{-3}$ |
| $S_{SiO_2}^{max}$ | SiO ₂ of most evolved composition ^{12,73,74} | 75 | 75 | % |
| $S_{SiO_2}^{min}$ | SiO ₂ of least evolved composition ^{12,73,74} | 50 | 50 | % |
| K | Trace element Nernst partition coefficient ¹⁶ | 0.08 | 0.08 | - |

Values used to produce the results shown in all figures except Extended Data Fig. 10. A steeper geotherm suitable for thermally mature crust²³ was assumed for the results shown in Extended Data Figs. 4–8, which have intrusion at 10 km depth. The range of values for the sensitivity analysis was used to calculate the range of values of the dimensionless scaling factor κ shown in Extended Data Fig. 10a and to produce the associated numerical modelling results shown in Extended Data Fig. 10b, c. Data sources^{12,16,21–26,32–34,42,50,51,57,58,73,74,76,77} are indicated.

Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes

Gordon Lax^{1,4}, Yana Eglit^{1,4}, Laura Eme^{2,3,4}, Erin M. Bertrand¹, Andrew J. Roger² & Alastair G. B. Simpson^{1*}

Almost all eukaryote life forms have now been placed within one of five to eight supra-kingdom-level groups using molecular phylogenetics^{1–4}. The ‘phylum’ Hemimastigophora is probably the most distinctive morphologically defined lineage that still awaits such a phylogenetic assignment. First observed in the nineteenth century, hemimastigotes are free-living predatory protists with two rows of flagella and a unique cell architecture^{5–7}; to our knowledge, no molecular sequence data or cultures are currently available for this group. Here we report phylogenomic analyses based on high-coverage, cultivation-independent transcriptomics that place Hemimastigophora outside of all established eukaryote supergroups. They instead comprise an independent supra-kingdom-level lineage that most likely forms a sister clade to the ‘Diaphoretickes’ half of eukaryote diversity (that is, the ‘stramenopiles, alveolates and Rhizaria’ supergroup (Sar), Archaeplastida and Cryptista, as well as other major groups). The previous ranking of Hemimastigophora as a phylum understates the evolutionary distinctiveness of this group, which has considerable importance for investigations into the deep-level evolutionary history of eukaryotic life—ranging from understanding the origins of fundamental cell systems to placing the root of the tree. We have also established the first culture of a hemimastigote (*Hemimastix kukwesjijk* sp. nov.), which will facilitate future genomic and cell-biological investigations into eukaryote evolution and the last eukaryotic common ancestor.

We identified two previously undescribed species of the rarely observed protist group Hemimastigophora (one *Spironema* and one *Hemimastix*) in enrichments from soil. Here we formally describe the newly identified *Hemimastix* species.

Hemimastix Foissner, Blatterer & Foissner 1988
Hemimastix kukwesjijk Eglit and Simpson, sp. nov.

Etymology. *Kukwesjijk* (approximate pronunciation, ‘ku-ga-wes-jij-k’). ‘*Kukwes-*’ (Mi’kmaq), a rapacious, hairy ogre from the traditions of the Mi’kmaq First Nation of Nova Scotia; ‘*-jijk*’, a diminutive plural suffix. ‘Little ogres’ reflects the predatory and hairy nature of this microorganism, and the use of Mi’kmaq language and tradition acknowledges the region in which the species was isolated.

Type material. The name-bearing hapantotype consists of trophic cells and dividing cells of strain BW2H that are osmium-fixed, sputter-coated and mounted for scanning electron microscopy. This material is deposited with the American Museum of Natural History (New York) with accession code AMNH_IJC 00267132. This material also contains prey *Spumella* sp. (Stramenopiles) and uncharacterized prokaryotes, both of which are explicitly excluded from the hapantotype.

Description. *Hemimastix* species, 16.5–20.5-μm long with 17–19 flagella per row.

Type locality. Bluff Wilderness Trail, Nova Scotia, Canada (44.6610154° N, 63.7674669° W); soil from mixed-species woodland.

Gene sequence. The partial small subunit ribosomal RNA (SSU rRNA) gene sequence of strain BW2H has been deposited in GenBank, accession code MF682191.

Comments. Cells are larger and have several more flagella than *Hemimastix amphikineta*, the only previously described species (14-μm by 7-μm cell body, 12 flagella per row⁶).

Cells of *H. kukwesjijk* are oval in profile with a blunt anterior projection (the capitulum) and two rows of flagella along their whole length (Fig. 1b, Extended Data Fig. 1). In cultivation as strain BW2H, live cells were 16.5–20.5-μm long by 7–12.5-μm wide ($18.3 \pm 1.1 \mu\text{m} \times 9.9 \pm 1.2 \mu\text{m}$; $n = 61$), with a sub-central, rounded nucleus and posterior contractile vacuole (Fig. 1c). Each row of 17–19 flagella (mean 18.4; $n = 25$) lay in a channel between the two thick thecal plates. The anteriormost 9 or 10 flagella were closely spaced, and the rest emerged from separate notches in the underlying plate (Fig. 1b, e). The capitulum was bordered by the overlapping anterior

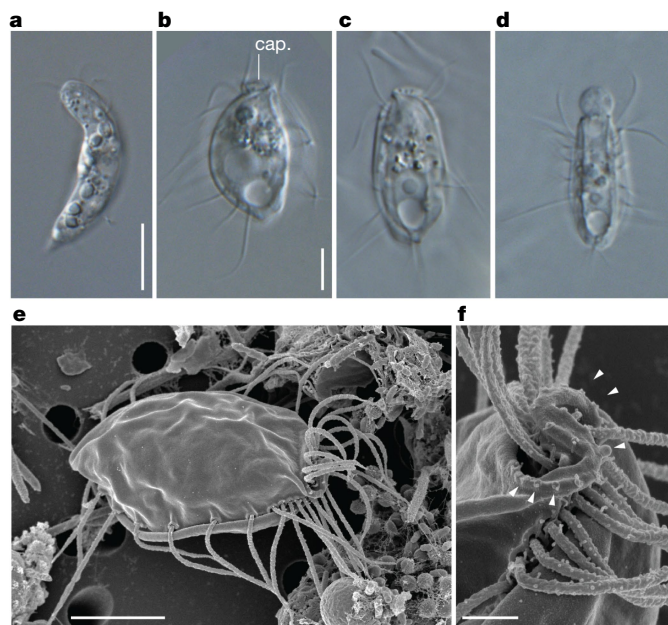


Fig. 1 | Micrographs of studied hemimastigotes. a, *Spironema* cf. *multiciliatum*, cell 1 (of 4) isolated for transcriptomics. **b–f**, *H. kukwesjijk*, cell 1 (of 2) isolated for transcriptomics (**b**); note the presence of the capitulum (cap.). **c**, **d**, Cells from culture (strain BW2H); note the nucleus and the contractile vacuole at the posterior (**c**), and feeding on prey with the capitulum (**d**). **e**, General view of cell (strain BW2H), anterior with the capitulum to right. **f**, Detail of the capitulum, showing caps of undischarged extrusomes (arrowheads) and close-spaced flagella in anterior part of flagellar rows. **a–d**, Differential interference contrast light microscopy. **e**, **f**, Scanning electron microscopy. Scale bars, 10 μm (**a**), 5 μm (**b–e**; scale bar in **b** applies to images **b–d**), 1 μm (**f**).

¹Department of Biology, Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, Nova Scotia, Canada. ²Centre for Comparative Genomics and Evolutionary Bioinformatics, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada. ³Present address: Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden. ⁴These authors contributed equally: Gordon Lax, Yana Eglit, Laura Eme. *e-mail: alastair.simpson@dal.ca

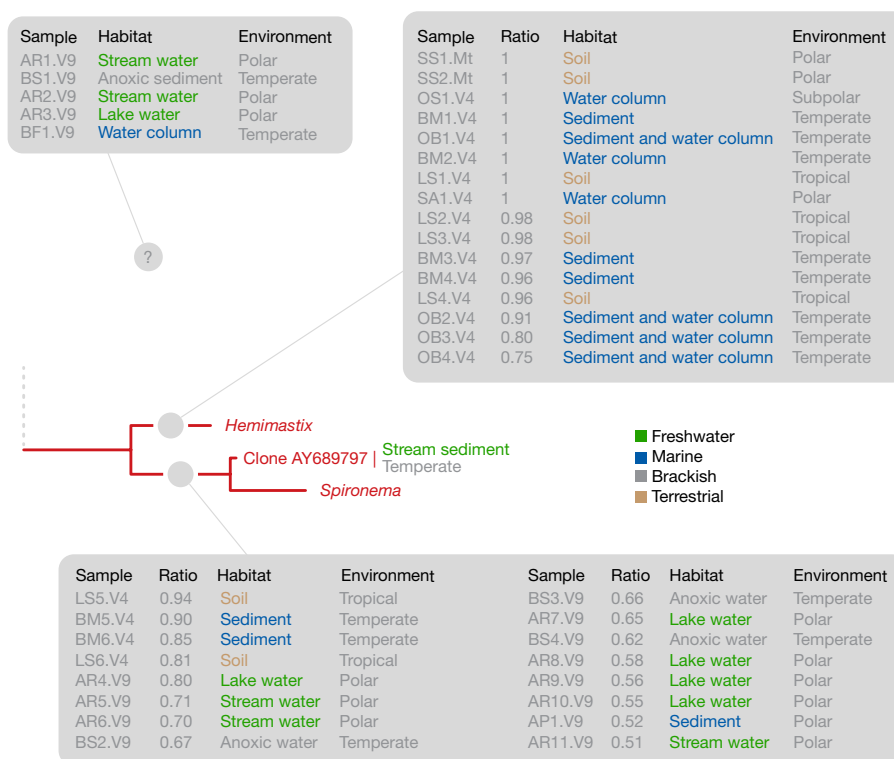


Fig. 2 | Environmental SSU rRNA/rDNA reads assigned to Hemimastigophora. The pplacer likelihood-to-weight ratio, habitat and environmental zone are reported for each read (denoted by sample code in the left columns). Reads with a likelihood-to-weight ratio > 0.5 are assigned to a branch. Five assigned sequences (denoted by a circled

?) were of uncertain placement within Hemimastigophora—that is, the likelihood-to-weight ratio for any single branch within the clade was < 0.5, but the sum of all likelihood-to-weight ratios = 1. See Extended Data Fig. 3 for full reference tree. Supplementary Table 1 gives additional information on individual reads, including sample codes.

ends of the flagellar rows, and the adjacent plate margins housed extrusomes (undischarged, Fig. 1f; discharged, Extended Data Fig. 2c). Cells fed on a small stramenopile (*Spumella* sp.) after attachment at the capitulum, and enclosure by the anterior flagella (Fig. 1d, Extended Data Figs. 1h–k, 2a, b).

The *Spironema* species awaits formal description; the cells we isolated—which we discuss here as *Spironema* cf. *multiciliatum* (see ‘Identification of *Spironema* cf. *multiciliatum*’ in Methods)—were spindle-shaped with a thin ‘tail’. These cells were 23–31-μm long by 4–7.5-μm wide (mean ± s.d., 27.4 ± 3.5 μm × 5.4 ± 1.6 μm; $n = 7$), with an oval nucleus and two rows of six or more flagella clustered in the anterior quarter, plus two or three flagella per row more posteriorly (Fig. 1a, Extended Data Fig. 1a).

We determined SSU rRNA sequences from both hemimastigotes, and used these to analyse published environmental sequence datasets to determine (1) the distribution of the group across habitats and (2) whether these sequences matched a known environmental clade. Unlike some other recently characterized lineages (for example, ref. ⁸), hemimastigotes do not appear to belong to a previously identified environmental clade. One unclassified long-read clone from freshwater sediment (AY689797) was phylogenetically related to *Spironema* (Fig. 2, Extended Data Fig. 3). An additional 37 short reads were detected among V4 or V9 amplicon datasets, or soil metatranscriptomes (Fig. 2, Supplementary Table 1). Many of the V4 and V9 amplicons derived from soil or freshwater, consistent with most light microscopy accounts⁷. However, nearly half of these amplicons came from marine sediment or water-column samples (Fig. 2), and one *Hemimastix*-like V4 amplicon was among the 25 most-abundant operational taxonomic units in a fjord sediment dataset (Supplementary Table 1).

To place hemimastigotes in the tree of eukaryotes, we generated transcriptomes from isolated single cells of both *Spironema* and *Hemimastix*, and assembled 351-gene datasets with a broad sampling of eukaryote taxa (initially 104 taxa; this was reduced to 61 taxa for

computationally intensive analyses). The transcriptomes proved to be high-coverage (*Spironema*, 290 of 351 = 82.6% of genes and 77.6% of sites represented; *Hemimastix*, 280 of 351 = 79.7% of genes, 72.1% of sites). Maximum likelihood analyses of both the 104-taxon and 61-taxon datasets were consistent with other recent phylogenomic studies^{3,9,10} in dividing previously known eukaryotes into three clans: Diaphoretickes, Discoba and an ‘Amorphea+’ assemblage (Fig. 3, Extended Data Fig. 4). The major subgroups of Diaphoretickes were Sar plus *Telonema*, Haptophyta plus Centrohelida, and Cryptista plus Archaeplastida and Picozoa. The ‘Amorphea+’ group contained Obazoa and Amoebozoa, as well as collodictyonids, rigifilids, *Mantamonas*, Ancyromonadida, Malawimonadida and Metamonada. The position of metamonads was unstable, which mirrors conflicts seen in other recent analyses^{9,11}.

Spironema and *Hemimastix* formed a maximally supported Hemimastigophora clade that was phylogenetically isolated. The 104-taxon analysis placed Hemimastigophora amongst the deepest branches within Diaphoretickes, as the sister of a clade of Sar, *Telonema*, haptophytes and centrohelids—though with equivocal support (ultra-fast bootstrap approximation = 83%; Fig. 4, Extended Data Fig. 4). In the 61-taxon analysis, Hemimastigophora again grouped with Diaphoretickes (bootstrap support = 100% (posterior mean site frequency method); ultra-fast bootstrap approximation = 93%; Bayesian posterior probability = 1), but actually branched sister to all of the other Diaphoretickes, which formed a clade (bootstrap support = 88%; ultra-fast bootstrap approximation = 60%; Bayesian posterior probability = 1; Fig. 3).

To further explore the position of Hemimastigophora, we analysed several derivatives of the 61-taxon dataset that excluded potential sources of phylogenetic inaccuracy. Analyses that (i) excluded the three taxa identified as outlier long-branches (dataset referred to as ‘58-nLB’), (ii) excluded the three data-poorest taxa (site coverage < 30%; dataset referred to as ‘58-nDP’) or (iii) recoded the amino

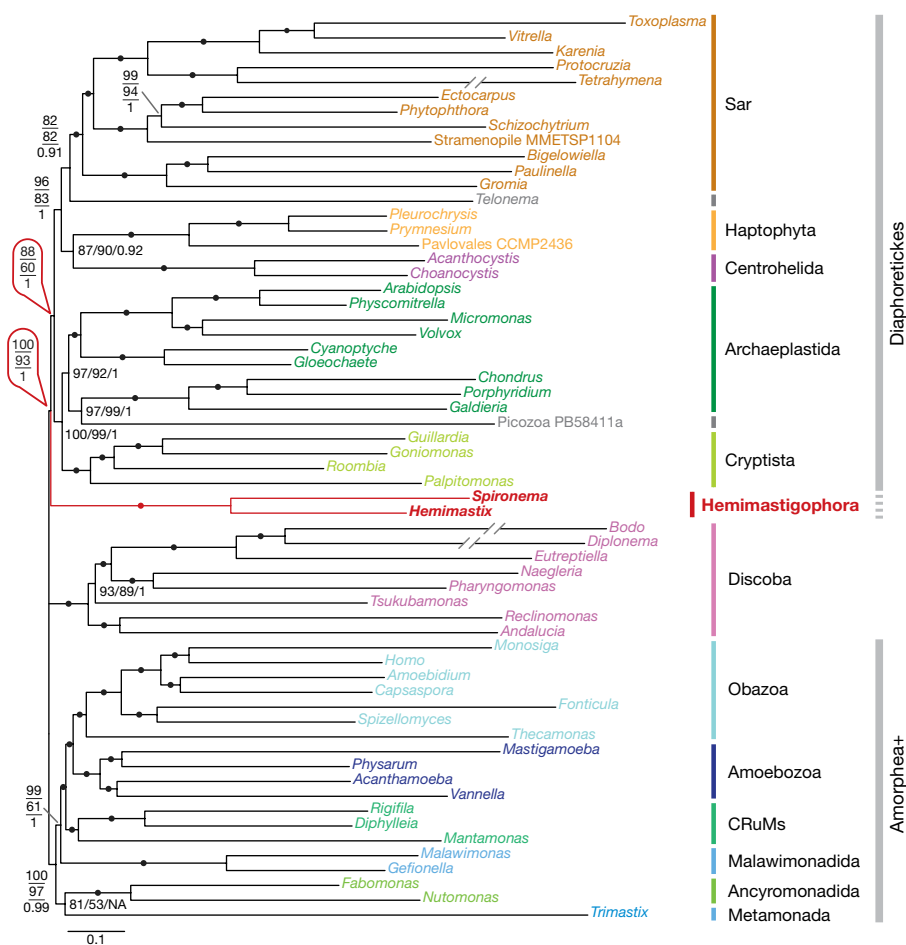


Fig. 3 | Phylogenetic placement of Hemimastigophora within eukaryotes. Unrooted phylogeny inferred from 351 genes and 61 taxa, using maximum likelihood under the 'LG + C60 + F + I' model. The numbers on branches show—in order from top to bottom or from left to right—posterior mean site frequency (PMSF) bootstrap percentages (bootstrap support; 200 true bootstrap replicates), ultrafast bootstrap approximation percentages (1,000 replicates) and Bayesian

posterior probabilities (under the 'CAT + GTR' model). Filled circles denote maximum support with all methods (that is, 100, 100 and 1, respectively). The three longest branches (leading to *Bodo*, *Diplonema* and *Tetrahymena*) are shown reduced by 1/3. CRuMs: collodictyonids, rigifilids and *Mantamonas*. NA, not available. Scale bar denotes 0.1 expected substitutions per site.

acid data into four categories (dataset referred to as '61-SR4') all supported the same topology as the original 61-taxa analysis—that is, Hemimastigophora outside of and sister to Diaphoretickes (Fig. 4, Extended Data Figs. 5–7). However, removing fast-evolving sites did not systematically favour the tree inferred in the 61-taxa analysis over a topology in which Hemimastigophora is sister to a Sar + *Telonema* + Haptophyta + Centrohelida clade (as in the 104-taxa analysis; Extended Data Fig. 8). Thus, although most analyses place Hemimastigophora as branching outside other Diaphoretickes, the alternative position—in which hemimastigotes fall one node inside Diaphoretickes—remains credible (Fig. 4).

All previous proposals for the phylogenetic or systematic placement of Hemimastigophora were based on morphology alone. The sub-membranous thecal plates between the two rows of flagella suggested an affinity with euglenids, which have a pellicle^{6,7}. Subsequently, affinities were proposed with completely different taxa that have pellicular or thecal structures—namely alveolates¹², or apusomonads and ancyromonads¹³. A placement within Rhizaria was also suggested on the basis of flagellum and extrusome substructure¹⁴. None of these proposals is supported by our phylogenies, because Hemimastigophora is always distantly related to euglenids (Euglenozoa, in Discoba), apusomonads and ancyromonads (both in Amorphea+) and Sar (which contains Alveolata and Rhizaria).

Instead, the extremely deep phylogenetic position of Hemimastigophora—most likely at the base of Diaphoretickes—implies

that they represent a novel, supra-kingdom-level lineage. This identifies hemimastigotes as a crucial group to include in descriptions of the tree of eukaryote life, and in most studies of the evolution of eukaryotic cells. This is especially important when inferring the history of eukaryotic innovations, or the nature of the last eukaryotic common ancestor, from the distributions across supergroups of particular genes, genome characteristics or cellular features^{15–19}. Hemimastigotes may be equally important in the immensely challenging task of placing the root of the eukaryote tree. The root is usually inferred^{20–23} to lie somewhere between the largest eukaryote clans—approximately in one of the positions marked a, b or c in Fig. 4—with position a (between Amorphea, and Diaphoretickes plus Discoba) currently being the most favoured^{22,23}. Hemimastigophora appears to lie close to all of these positions on the unrooted tree (see Fig. 4), and could be our only known representative of one of the most ancient divisions amongst extant eukaryotes. Accordingly, we searched the single-cell transcriptomes for genes that could have arisen during the divergences between supergroups (Fig. 4, Supplementary Table 3). We found several genes in hemimastigotes that are not known from Diaphoretickes, including those for myosin II—previously known from Amorphea, and one subgroup of Discoba^{18,24}—and Golgi protein GCP16 (also known as golgin A7) (previously specific to Amorphea)¹⁹. The presence of such genes in hemimastigotes either pushes back their likely origins to before the last eukaryotic common ancestor (or supports this inference) or—more controversially—could be due to the root of

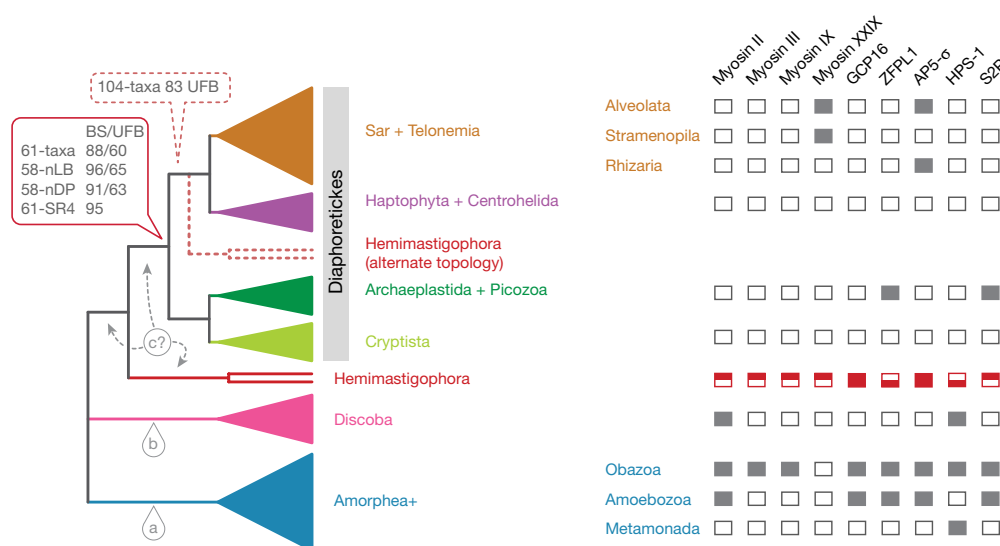


Fig. 4 | Summary of phylogenomic analyses and distribution of select genes across eukaryotes. Left, inferred phylogenetic positions of Hemimastigophora. Box with solid outline details the support for Hemimastigophora as a deep branch relative to the ‘Diaphoretickes’ supergroup, in various analyses. BS, PMSF bootstrap support (except 61-SR4 for which the ‘GTR + R6 + F’ model was used). UFB, ultrafast bootstrap approximation support. Dashed box shows support for the alternative topology (Hemimastigophora as a deep branch within Diaphoretickes) in the 104-taxa analysis. Stepwise fast-site removal

eukaryotes being further from the base of Amorphea than generally supposed²³—that is, Amorphea and Hemimastigophora being on the same side of the root (shown by the top variant of position c in Fig. 4). However, another hemimastigote myosin-family gene was previously unknown outside the Sar clade (Fig. 4): irrespective of the final position of the root, this survey demonstrates that the antiquity of gene origins tends to be underestimated until all major lineages are considered. This bias can result in the underestimation of the gene content of ancient eukaryotes, and thus overestimations of the simplicity of their cell biology. Examining hemimastigote genomes—and ultimately their cell biology—will be valuable for better understanding eukaryote evolution at the deepest levels.

This study has used single-cell transcriptomics to unveil a deep-branching eukaryote lineage. Single-cell transcriptomics and genomics^{25–27} bypass the ‘culture bottleneck’ and thus provide a rapid path to deeper taxon sampling, even when species from a group of interest are eventually cultivated. This is particularly valuable for phylogenomics, in which inaccuracy owing to poor taxon sampling is a perpetual concern²⁸. For this application, single-cell transcriptomics outperforms single-cell genomics because of better coverage of house-keeping genes (see, for example, refs ^{26,27}). Information on multiple related species is also valuable for ensuring data fidelity (detecting contaminants, gene transfers and so on; see Methods). Single-cell techniques are especially promising for the heterotrophic protozoa that probably represent most ‘undiscovered’ major lineages, and for which establishing cultures with suitable prey or hosts can be challenging^{25,27,29,30}.

In this molecular phylogenetic investigation of Hemimastigophora, we show that they are a previously unrecognized supergroup of eukaryotes. Their phylogenetic distinctiveness is comparable to the whole animal plus fungi clade (Opisthokonta) or the assemblage containing all land plants and primary algae (Archaeplastida). We expect the discovery or recognition of other important lineages will greatly accelerate owing to similar applications of single-cell methods.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0708-8>.

analyses equivocated between these alternatives (see Extended Data Fig. 8). Labels a, b and c show the possible positions of the eukaryote root; the likely placement of Hemimastigophora results in several variants of position c. Right, known distributions of selected proteins encoded by genes with proposed deep origins among living eukaryotes that were detected in hemimastigote transcriptomes. Boxes filled in their top half denote genes detected in *Spironema*; boxes filled in their bottom half denote genes detected in *Hemimastix*; completely filled boxes represent genes detected in both organisms; see Supplementary Table 3 for details.

Received: 26 October 2017; Accepted: 21 September 2018;
Published online 14 November 2018.

- Burki, F. The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harb. Perspect. Biol.* **6**, a016147 (2014).
- Worden, A. Z. et al. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. *Science* **347**, 1257594 (2015).
- Burki, F. et al. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. *Proc. R. Soc. Lond. B* **283**, 20152802 (2016).
- Simpson, A. G. B. & Eglit, Y. in *Encyclopedia of Evolutionary Biology* Vol. 3 (ed. Kliman, R. M.) 344–360 (Elsevier, Amsterdam, 2016).
- Klebs, G. *Flagellatenstudien* (Akademische Verlags-Gesellschaft, Leipzig, 1893).
- Foissner, W., Blatterer, H. & Foissner, I. The Hemimastigophora (*Hemimastix amphikineta* nov. gen., nov. spec.), a new protistan phylum from Gondwanian soils. *Eur. J. Protistol.* **23**, 361–383 (1988).
- Foissner, I. & Foissner, W. Revision of the family Spironemidae Doflein (Protista, Hemimastigophora), with description of two new species, *Spironema terricola* n. sp. and *Stereonema geiseri* n. g., n. sp. *J. Eukaryot. Microbiol.* **40**, 422–438 (1993).
- Yubuki, N. et al. Morphological identities of two different marine stramenopile environmental sequence clades: *Bicosoeca kenaiensis* (Hilliard, 1971) and *Cantina marsupialis* (Larsen and Patterson, 1990) gen. nov., comb. nov. *J. Eukaryot. Microbiol.* **62**, 532–542 (2015).
- Brown, M. W. et al. Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. *Proc. R. Soc. Lond. B* **280**, 20131755 (2013).
- Zhao, S. et al. *Collodictyon*—an ancient lineage in the tree of eukaryotes. *Mol. Biol. Evol.* **29**, 1557–1568 (2012).
- Cavalier-Smith, T. et al. Multigene eukaryote phylogeny reveals the likely protozoan ancestors of opisthokonts (animals, fungi, choanozoans) and Amoebozoa. *Mol. Phylogenet. Evol.* **81**, 71–85 (2014).
- Cavalier-Smith, T. A revised six-kingdom system of life. *Biol. Rev. Camb. Philos. Soc.* **73**, 203–266 (1998).
- Cavalier-Smith, T. in *The Flagellates, The Systematics Association Special Volume Series 59* (eds Leadbeater, B. S. C. & Green, J. C.) 361–390 (Taylor & Francis, London, 2000).
- Cavalier-Smith, T., Lewis, R., Chao, E. E., Oates, B. & Bass, D. Morphology and phylogeny of *Sainouron acronematica* sp. n. and the ultrastructural unity of Cercozoa. *Protist* **159**, 591–620 (2008).
- Speijer, D., Lukeš, J. & Eliáš, M. Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life. *Proc. Natl Acad. Sci. USA* **112**, 8827–8834 (2015).
- de Mendoza, A. et al. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proc. Natl Acad. Sci. USA* **110**, E4858–E4866 (2013).
- Fukasawa, Y., Oda, T., Tomii, K. & Imai, K. Origin and evolutionary alteration of the mitochondrial import system in eukaryotic lineages. *Mol. Biol. Evol.* **34**, 1574–1586 (2017).

18. Seb  -Pedr  s, A., Grau-Bov  , X., Richards, T. A. & Ruiz-Trillo, I. Evolution and classification of myosins, a paneukaryotic whole-genome approach. *Genome Biol. Evol.* **6**, 290–305 (2014).
19. Barlow, L. D., N  ylvtov  , E., Aguilar, M., Tachezy, J. & Dacks, J. B. A sophisticated, differentiated Golgi in the ancestor of eukaryotes. *BMC Biol.* **16**, 27 (2018).
20. He, D. et al. An alternative root for the eukaryote tree of life. *Curr. Biol.* **24**, 465–470 (2014).
21. Katz, L. A., Grant, J. R., Parfrey, L. W. & Burleigh, J. G. Turning the crown upside down: gene tree parsimony roots the eukaryotic tree of life. *Syst. Biol.* **61**, 653–660 (2012).
22. Derelle, R. & Lang, B. F. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Mol. Biol. Evol.* **29**, 1277–1289 (2012).
23. Derelle, R. et al. Bacterial proteins pinpoint a single eukaryotic root. *Proc. Natl Acad. Sci. USA* **112**, E693–E699 (2015).
24. Richards, T. A. & Cavalier-Smith, T. Myosin domain evolution and the primary divergence of eukaryotes. *Nature* **436**, 1113–1118 (2005).
25. Kolisko, M., Boscaro, V., Burki, F., Lynn, D. H. & Keeling, P. J. Single-cell transcriptomics for microbial eukaryotes. *Curr. Biol.* **24**, R1081–R1082 (2014).
26. Yoon, H. S. et al. Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* **332**, 714–717 (2011).
27. Gawryluk, R. M. R. et al. Morphological identification and single-cell genomics of marine diplomonads. *Curr. Biol.* **26**, 3053–3059 (2016).
28. Keeling, P. J. et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.* **12**, e1001889 (2014).
29. Caron, D. A. et al. Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nat. Rev. Microbiol.* **15**, 6–20 (2017).
30. Kr  bber  d, A. K. et al. Single cell transcriptomics, mega-phylogeny, and the genetic basis of morphological innovations in Rhizaria. *Mol. Biol. Evol.* **34**, 1557–1573 (2017).

Acknowledgements The authors thank P. Li and P. Scallion (Dalhousie University) for assistance with electron microscopy, M. Dlutek (Dalhousie

University) for Illumina sequencing, S. Geisen (Wageningen University) for providing parsed metatranscriptomic data, F. Mah   (CIRAD, Montpellier) for access to and parsing much of the V4 data, M. Brown (Mississippi State) for the seed phylogenomic dataset, A. Seb  -Pedr  s (Weizmann Institute of Science) for the seed myosin alignments, M. Kolisko (Institute of Parasitology, Czech Academy of Sciences) for data handling scripts, B. Q. Minh (University of Vienna) for substantial help with phylogenomic analyses and troubleshooting in IQ-TREE, and R. Lewis (Nova Scotia Museum) and B. Francis for advice on MiSeq tradition and language. This work was supported by CIFAR, NSERC grant 298366-2014 to A.G.B.S. and NSERC grant 2016-016792 to A.J.R.

Reviewer information Nature thanks I. Ruiz-Trillo and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.E. isolated the organisms and cultivated *H. kukwesjijk*. Y.E. and G.L. undertook the microscopy. G.L. performed the single-cell transcriptomics. Y.E., G.L. and E.M.B. analysed the rDNA and environmental sequence data. G.L., L.E., Y.E. and A.G.B.S. assembled the phylogenomic datasets. G.L., L.E. and A.J.R. performed phylogenomic analyses. L.E. and Y.E. performed the gene presence analyses. G.L., Y.E. and A.G.B.S. wrote the manuscript, with input from all co-authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0708-8>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0708-8>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to A.G.B.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size.

Cell isolation and transcriptomics. Soil from mixed-species woodland in the Bluff Wilderness Trail in Nova Scotia, Canada (44.6610154° N, 63.7674669° W; 17 April 2016) was kept hydrated with distilled water in a Petri dish until hemimastigotes were observed about four weeks later. Single *Spironema* and *Hemimastix* cells were isolated with drawn-out micropipettes, photo-documented by differential interference contrast light microscopy (using a Zeiss Axiovert 200M and AxioCam ICc5 microscope and camera system; Carl Zeiss AG), and subjected to single-cell transcriptomics using the Smart-seq2 protocol³¹ with modifications. In brief, four (*Spironema*) or two (*Hemimastix*) cells were individually picked into 0.2% Triton X-100 lysis buffer, immediately frozen in liquid nitrogen, then thawed and re-frozen three times. The remaining procedure followed the original protocol, with 20 (*Spironema*) or 18 (*Hemimastix*) PCR cycles. cDNA quantity and quality was assessed (i) by Qubit dsDNA HS assay (Thermo Fisher, Q32851) and (ii) by PCR, and cloning of cDNA fragments into StrataClone SoloPack competent cells (Agilent Technologies), and 12 clones each were Sanger-sequenced. After library preparation with Illumina Nextera XT, sequencing was carried out on an Illumina MiSeq with 2 × 250-bp dual reads, with the libraries multiplexed on the same run.

Cultivation of *H. kukwesjijk*. To cultivate *H. kukwesjijk* strain BW2H, three cells were picked and washed with a micropipette, then transferred to a prey—*Spumella* sp. (strain BW2S)—that was cultured by serial dilution from the same sample. Cultures were maintained in 15-ml tubes containing ~4 ml of 25%-strength ATCC medium 802, with one sterilized barley grain, angled for aeration and transferred weekly. Cells were examined by light microscopy as described above.

Scanning electron microscopy of *H. kukwesjijk*. Cells from a ten-day-old culture of strain BW2H were fixed for 30 min in OsO₄ vapour alone (at room temperature) or OsO₄ vapour simultaneously with 2.5% glutaraldehyde (on ice), and filtered onto 2-µm isopore membrane filters (Millipore). These were washed in distilled water and dehydrated in a series of 50–100% ethanol mixtures, critical-point-dried in CO₂ and sputter-coated by 10 nm of gold-palladium. Cells were imaged using a Hitachi S-4700 SEM at 3 kV.

SSU rDNA analyses. A single cell of *Spironema* cf. *multiciliatum* was isolated and washed by micropipetting, and then photo-documented (see above). The genomic DNA of this cell was amplified using multiple displacement amplification (Illustra GenomiPhi V3 DNA amplification kit, GE Healthcare). Total genomic DNA was extracted from *H. kukwesjijk* culture BW2H (also including the prey *Spumella* sp., strain BW2S) using a Qiagen DNeasy kit. Partial SSU rDNA sequences were PCR-amplified from *Spironema* cf. *multiciliatum* and *Spumella* sp. BW2S using primers 82F (5'-GAAACTGCGAATGGCTC-3') and 1498R (5'-CACCTACGGGAACCTTGTTA-3'), with annealing temperatures of 58 °C and 55 °C, respectively. A partial *Hemimastix* SSU rDNA sequence was PCR-amplified from strain BW2H using exact-match primers Hemi2-342F (5'-ACTTTCGATTGTAGGATAGA-3') and Hemi2-1103R (5'-AAAACCTGCGATTCTCTGG-3') with an annealing temperature of 55 °C. All amplicons were directly Sanger-sequenced at Génome Québec. The SSU rDNA of *Spumella* sp. strain BW2S was 99% identical to *Spumella* strain 187hm (GenBank accession code: DQ388550).

The SSU rRNA sequences for the two hemimastigotes were extracted from the transcriptome data (see above) and compared to the SSU rDNA sequences obtained independently from genomic DNA, to ensure mutual identity (although the rDNA sequence of *H. kukwesjijk* did differ from the transcriptome-derived rRNA sequence in having a 395-bp intron). The transcriptome-derived SSU rRNA sequences (and environmental clone AY689797, retrieved from GenBank via megablast) were then added via profile alignment using MUSCLE³² to a global eukaryotic alignment of SSU rRNA genes (111 taxa total). Following manual inspection of the alignment, poorly aligned sites were masked using Gblocks³³ with subsequent manual correction (1,252 sites retained), and a phylogeny was estimated in RAxML under the 'GTR + Γ ' model³⁴ with a 1,000-replicate bootstrap analysis (Extended Data Fig. 3).

Environmental SSU rRNA and rDNA sequence comparisons. Sequences derived from eukaryotic environmental SSU rRNA and rDNA were acquired from VAMPS³⁵ (V9), TARA Oceans³⁶ (V9), BioMarKs³⁷ (V4), a neotropical soil study³⁸ (V4), a high-arctic Fjord water column study³⁹ (V4) and a soil metatranscriptome dataset⁴⁰, and queried in a BLAST⁴¹ analysis with the appropriate (V4 or V9) section of the *Spironema* and *Hemimastix* SSU rRNAs, at a 85% identity cut-off (top 500 hits). The corresponding short reads from the datasets were first aligned to the eukaryote reference alignment (see above) using PaPaRa⁴² version 2.5 and then placed on the SSU rRNA gene tree (Extended Data Fig. 3) using pplacer⁴³ version 1.1. Chimeric reads were identified manually with BLAST against the Genbank non-redundant nucleotide database and discarded (all cases were from VAMPS V9 datasets). Reads were also discarded if the top 100 BLAST hits were all to a single taxonomic group (for example, ciliates). Surviving reads were assigned to Hemimastigophora if they were placed on a particular branch within

Hemimastigophora with a likelihood-to-weight ratio > 0.5, or if they had an accumulated likelihood-to-weight ratio > 0.9 across the multiple branches within Hemimastigophora.

Phylogenomic dataset assembly. To perform phylogenomic analyses of eukaryotes that included hemimastigotes, we used the single-cell transcriptomes derived from *Hemimastix* (2 cells) and *Spironema* (4 cells), as described above. Raw reads from the Illumina sequencing were quality-trimmed, and the adapters clipped with Trimmomatic version 0.32⁴⁴ (default parameters), then assembled with Trinity⁴⁵ version 2.0.2 (default parameters). Assemblies were cleaned of sequencing cross-contamination using a custom script. Marker genes of interest were extracted using a previously reported pipeline⁴⁶ and appended as translated peptide sequences to a 396-taxon, 351-gene eukaryote dataset⁴⁶. This dataset was pruned to 107 taxa that broadly represented all major eukaryotic groups for which data were available, while excluding extremely 'long-branching' species and—where possible—species with poor sampling of this gene set. The 351 single-gene dataset was aligned individually using MAFFT-L-INS-i⁴⁷ version 7.0, and trimmed with BMGE⁴⁸ version 1.0 (-m BLOSUM30 -h 0.5 -g 0.2). From the resulting files, single-gene trees were generated with IQ-TREE⁴⁹ version 1.4.4 under the LG + C20 + F + G model with a 1,000-replicate, ultra-fast bootstrap approximation to estimate branch support⁵⁰. These trees were manually checked for sequences corresponding to probable paralogs, contaminants, or lateral- or endosymbiotic gene transfers, which were then removed from the datasets. The tree estimation and manual checking was then repeated, and any additional suspect sequences removed. Three taxa with limited remaining data (< 10% of sites) were then excluded, leaving 104 taxa for initial phylogenomic analysis.

Quality of hemimastigote transcriptomes. It was particularly important to assess the quality of the data from *Spironema* and *Hemimastix*, both because they were the subject of the study and because they were derived using single-cell methods from crude enrichments. The transcriptome from *Spironema* included 290 of the 351 genes in the phylogenomic dataset (82.6%) and 77.6% of the sites retained after trimming. The transcriptome from *Hemimastix* included 280 out of 351 = 79.7% of genes, and 72.1% of sites. In other words, both transcriptomes were reasonably data-rich from a phylogenomic perspective, and compare well to many transcriptomes from cultivated non-model protists (Supplementary Table 2). In all, 247 of the 351 gene alignments (70.4%) included both taxa. The *Spironema* and *Hemimastix* sequences formed a clade in 168 of the 247 (68%) single-gene trees inferred for these data, which is consistent with a specific relationship between the two hemimastigotes, bearing in mind that some of the individual genes in the dataset carry relatively little phylogenetic signal. There was no particular pattern to the relationships between each hemimastigote and other eukaryotes in the remaining 32% of trees. In summary, the single-gene trees indicate that there was little-to-no contamination from other eukaryotes in the analysed hemimastigote data. Furthermore, the *Spironema* and *Hemimastix* sequences always differed in these 247 alignments, which confirms that no cross-contamination between the two had carried through to the final dataset.

Phylogenomic analyses. The 351 individual-gene alignments with 104 retained taxa (see above) were concatenated, and trimmed with BMGE (-m BLOSUM30 -h 0.42 -g 1), yielding a dataset that consisted of 104 taxa and 93,798 amino acid sites. To enable more-complex analyses, we then excluded 43 phylogenetically redundant taxa—followed by re-trimming with BMGE (as above)—to generate a 61-taxon dataset with 93,903 amino acid sites. Taxa were selected for retention in the 61-taxon dataset such that eukaryote diversity remained reasonably evenly sampled, and that all major taxa that were included in the 104-taxon dataset were still represented. Where there was a choice, species with high gene coverage were retained in preference to species that were more poorly sampled, and shorter-branching species were retained over longer-branching species. Phylogenies for both datasets were inferred by maximum likelihood using IQ-TREE under the LG + C60 + F + Γ mixture model, with robustness assessed by ultra-fast bootstrap approximation (1,000 replicates). The 61-taxon dataset was also subjected to a 'full' bootstrap analysis with 200 replicates under the PMSF model, implemented in IQ-TREE. PMSF is a site-heterogeneous mixture model that can closely approximate complex mixture models such as LG + C60 + F + Γ while reducing computational time several-fold⁵¹, making full bootstrapping practical for our ~60-taxon datasets. The maximum likelihood tree that was inferred for this dataset under the LG + C60 + F + Γ model (see above) was used as the guide tree for the PMSF analysis. The 61-taxon dataset was also subjected to Bayesian analysis with PhyloBayes⁵² version 4.1 under the CAT + GTR model⁵³, with default priors and Markov chain Monte Carlo settings. Four independent Markov chain Monte Carlo chains were run for ~10,000 generations. Three chains converged (maximum difference in posterior probability < 0.13; burn-in = 3,000). Their consensus tree shows Hemimastigophora as sister to (other) Diaphoretickes with maximal support (that is, consistent with the maximum likelihood tree), whereas the unconverged chain yielded the topology in which Hemimastigophora is sister to the Sar + Telonema + Haptophyta + Centrohelida grouping.

Several further sets of analyses were conducted on derivatives of the 61-taxon dataset. First, we used a custom script to calculate average tip-to-tip distances for each taxon and identify 'long-branching' outliers (that is, taxa for which the average tip-to-tip branch lengths were longer than three standard deviations from the centre of the distribution of average branch lengths). Removing the three identified outliers (*Bodo*, *Diplonema* and *Tetrahymena*) yielded the '58 taxa, no long-est branches' (58-nLB) dataset. This was analysed using maximum likelihood, as per the main 61-taxon analysis (IQ-TREE with LG + C60 + F + Γ , with 1,000-replicate ultra-fast bootstrap approximation, and 200 bootstraps using PMSF with the LG + C60 + F + Γ maximum likelihood tree for the 58-nLB dataset as the guide tree).

Second, we deleted the three most data-poor taxa, each of which had site coverage < 30% (*Telonema*, *Gromia* and the picozoan PB58411a), resulting in a '58 taxa, no data-poor species' (58-nDP) dataset. This was analysed using maximum likelihood as per the main 61-taxon analysis, except that the PMSF bootstrap analysis was based on 100 replicates.

Third, we recoded the main 61-taxon dataset into four distinct categories of amino acids (SR4 scheme⁵⁴), to address possible compositional heterogeneity. The resulting 61-SR4 dataset was analysed with IQ-TREE under a GTR + R6 + F model, with 500 real bootstrap replicates.

Fourth, we used the assignment of per-site rates in IQ-TREE (-wsr flag) for the main 61-taxon dataset, and progressively removed the fastest-evolving sites in 10 steps, with approximately 4% of the sites removed in each step. This yielded 10 'stepwise fastest sites removed' (61-SFSR) datasets. To exclude the influence of the position of Hemimastigophora in the guide trees for subsequent PMSF analyses, we deleted the two hemimastigotes from the full dataset and the 10 SFSR datasets (that is, 11 total) with phyx version 0.1⁵⁵, and pruned these two species from the maximum likelihood tree from the 61-taxon dataset. The pruned tree was then used as the guide tree to calculate PMSF profiles ('PMSF-nHEMI') under LG + C60 + F + Γ . For each of the original 11 datasets (that is, datasets that included hemimastigotes), we then inferred support for important bipartitions under this LG + C60 + F + Γ PMSF model using a 1,000-replicate, ultra-fast bootstrap approximation, and plotted these support values against the percentage of sites remaining (Extended Data Fig. 8). This method of generating the PMSF model (PMSF-nHEMI) and evaluating statistical support differs from the main analyses (for example, 61-taxon, 58-nLB or 58-nDP), and the support values cannot be directly compared between these analyses and the 61-SFSR analyses.

Identification of non-universal ancient genes. To search the hemimastigote transcriptome data for gene innovations that potentially originated early in the evolution of crown eukaryotes (and thus may also represent synapomorphies that provide information about the relationships between supergroups), we collated a set of gene systems reported in the literature to include genes with widespread—but not universal—distributions across major eukaryote groups. Specific genes were selected on the basis of their presence in more than one species-rich 'supergroup' of eukaryotes—for example both Obazoa and Amoebozoa (see Supplementary Table 3). For this purpose, Metamonada and Discoba were considered distinct supergroups. Sequences were retrieved from GenBank or from the literature, and used as BLASTp queries against both hemimastigote transcriptomes, translated into amino acid sequences using a custom script (default genetic code). Where genes were not identified with BLASTp, hidden Markov model profiles were obtained either from the PFAM database or the literature (as indicated in Supplementary Table 3), or were built de novo from the alignments in the corresponding literature using hmmbuild, and then scanned for in both hemimastigote transcriptomes using hmmsearch (both hmmbuild and hmmsearch from the Hmmer-3.1b2 package⁵⁶). Genes that were retrieved in only one of the hemimastigote transcriptomes were used as BLASTp queries against the other. Hemimastigote candidate orthologues were verified by reciprocal BLASTp against the nr database, and—where appropriate—domain annotation databases (InterProScan and SMART), and then added to pre-existing alignments from corresponding references (as shown in Supplementary Table 3) via profile alignment using MUSCLE in Seaview version 4.6^{32,57}. Where phylogenies were necessary to further confirm identity (particularly in the case of multigene families), the alignments were trimmed using BMGE version 1.1⁴⁸ (-m BLOSUM30), and phylogenies estimated in IQ-TREE version 1.5.49 under the LG4X model. An alignment for HPS1 was not available in the original publication and was instead assembled from sequences from GenBank and publicly available transcriptomes, and aligned via MAFFT-L-INS-i⁴⁷. Because of the large size of the myosin gene family and the level of divergence between various paralogues, myosin homologues were instead aligned with MAFFT-E-INS-i and trimmed less conservatively (BMGE; -m BLOSUM30 -b 2), with the corresponding phylogeny estimated under the LG + C60 + F + Γ model.

Identification of *Spironema* cf. *multiciliatum*. The cells we discuss as *Spironema* cf. *multiciliatum* have an elongate shape (Fig. 1a, Extended Data Fig. 1a) and the 'main row' of flagella is restricted to the anterior portion. These features identify this organism with *Spironema* rather than *Hemimastix* (broad and flattened),

Paramastix (globular) or *Stereonema* (elongate but main rows of flagella about half the length of the cell^{7,58,59}). There are three previously described species of *Spironema*: *Spironema terricola*, *Spironema goodeyi* and *Spironema multiciliatum*. The shape and size of our specimens is inconsistent with *S. terricola* and *S. goodeyi*, both of which are very long and thin⁷. In addition, neither of these species has any posterior flagella. Our cells are similar in shape to *S. multiciliatum*⁵. The number of flagella in the 'main row' and the presence of a few difficult-to-observe flagella towards the posterior end are also broadly consistent with a previous account of *S. multiciliatum*, in which such posterior flagella were seen in some cells^{5,7}. However, our cells are 23–31 μm in length (mean: 27.4 μm (s.d., 3.45 μm); $n = 7$; see main text), which is markedly longer than the 18- μm length reported for *S. multiciliatum*. Thus, we determined that our specimens are similar—but not identical—to *S. multiciliatum*.

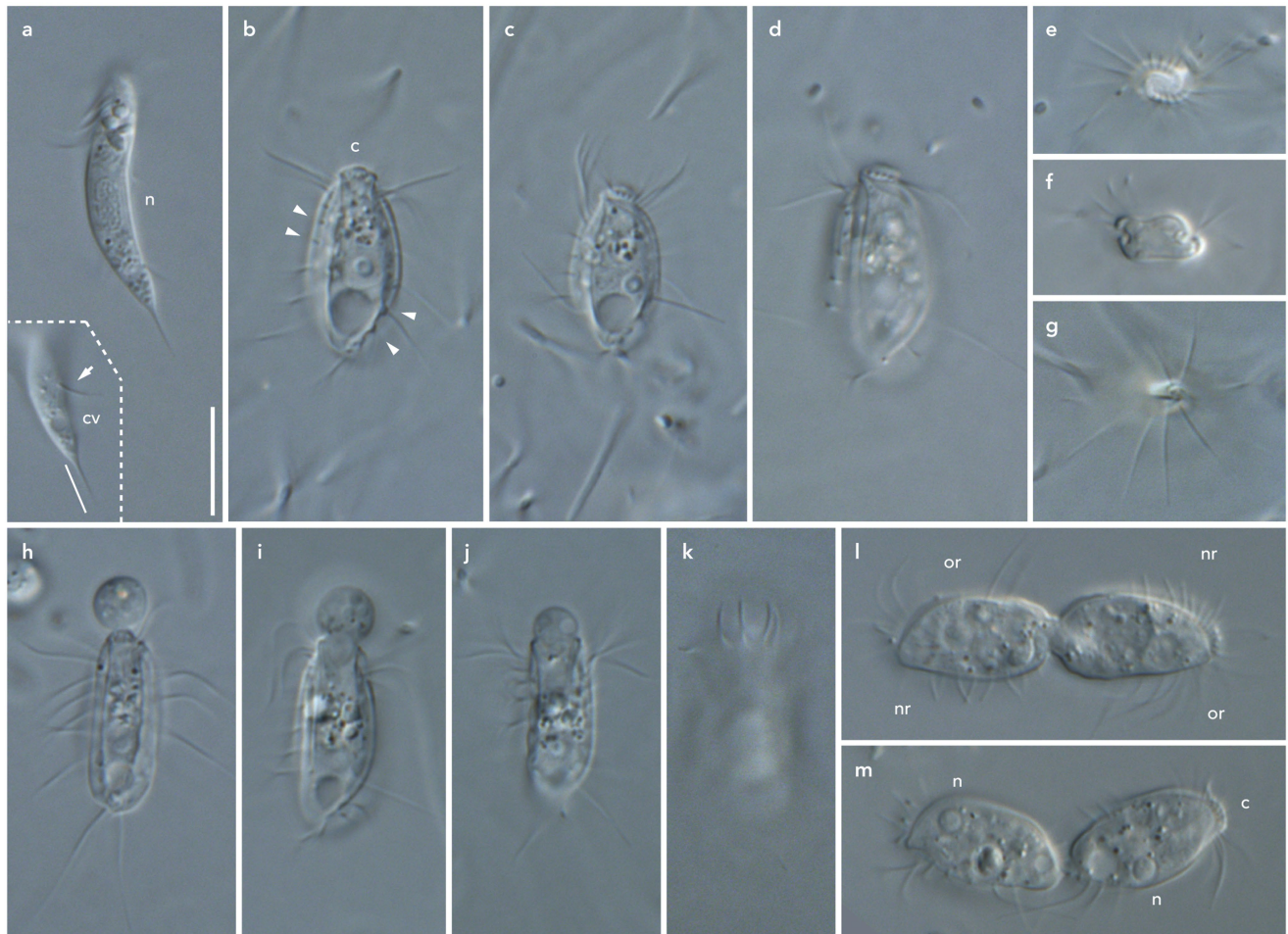
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Raw reads of *Spironema* and *Hemimastix* transcriptomes are deposited in GenBank under accession codes SRR6032743 and SRR6032744, respectively. The assembled *Hemimastix* and *Spironema* transcriptomes, 351 individual-gene alignments (104 taxa), concatenated and trimmed alignments and tree-files for the 104-taxon, 61-taxon, 58-nLB, 58-nDP, 61-SR4 and 61-SFSR datasets, alignments and tree files for non-universal ancient genes, raw light microscopy and scanning electron microscopy images, and the SSU rDNA alignment and tree-files have been deposited in Dryad (<https://doi.org/10.5061/dryad.n5g39d7>). The partial SSU rDNA gene sequence of *H. kukwesjijk* strain BW2H is deposited in GenBank, under accession code MF682191. This publication has been registered with the ZooBank database (<http://zoobank.org/>) with the Life Science Identifier urn:lsid:zoobank.org:pub:4BA2A83C-8363-4EBE-A9C7-097CA470F9FB, and the name *Hemimastix kukwesjijk* has been deposited in ZooBank with the Life Science Identifier urn:lsid:zoobank.org:act:32E12332-A418-40E2-BF4C-F2BFD94BF4CF.

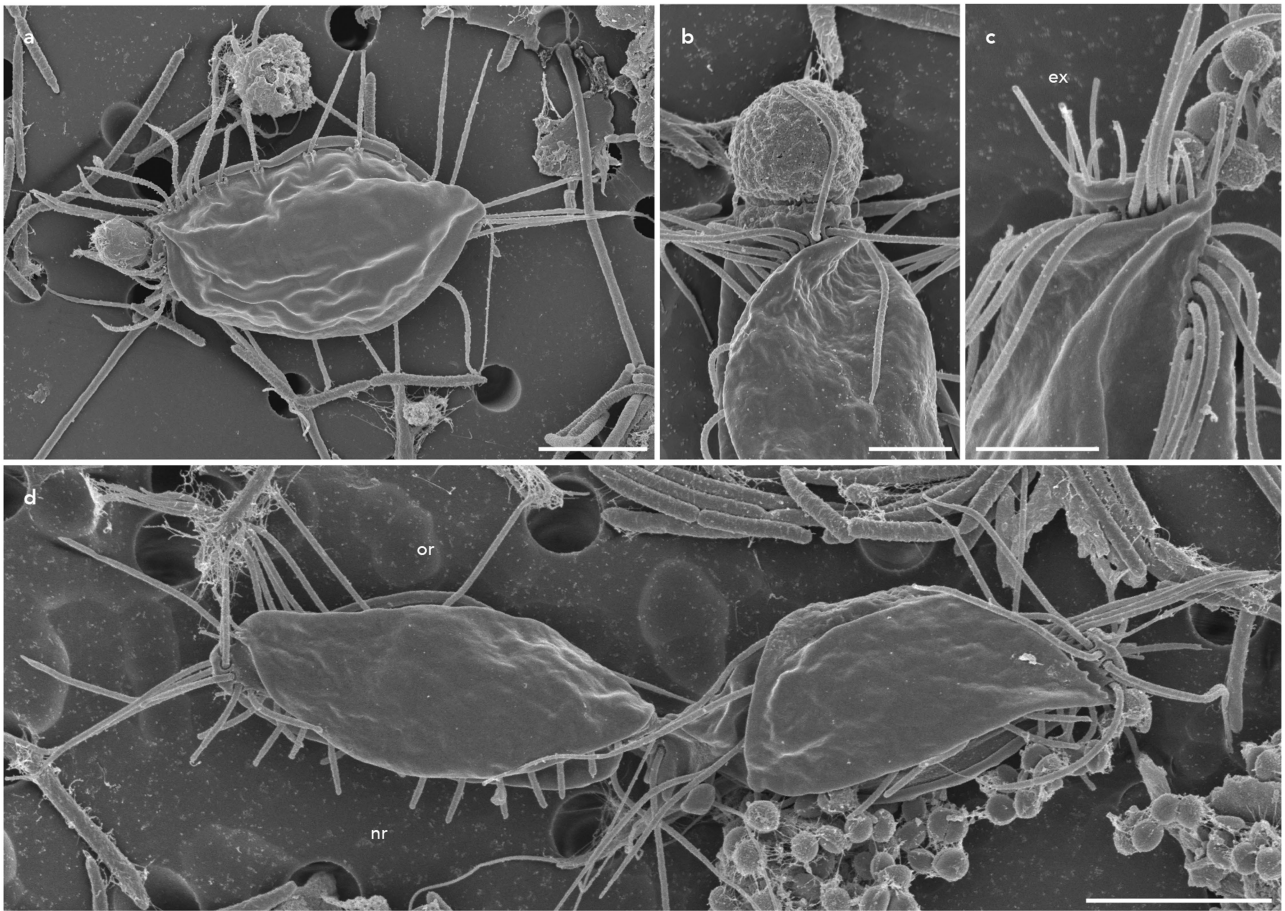
- Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
- Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
- Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
- Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
- Huse, S. M. et al. VAMPS: a website for visualization and analysis of microbial population structures. *BMC Bioinformatics* **15**, 41 (2014).
- de Vargas, C. et al. Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).
- BioMarks Consortium. BioMarks data portal <http://www.biomarks.eu> (2011).
- Mahé, F. et al. Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nat. Ecol. Evol.* **1**, 0091 (2017).
- Marquardt, M., Vader, A., Stübner, E. I., Reigstad, M. & Gabrielsen, T. M. Strong seasonality of marine microbial eukaryotes in a high-arctic fjord (Isfjorden, in West Spitsbergen, Norway). *Appl. Environ. Microbiol.* **82**, 1868–1880 (2016).
- Geisen, S. et al. Metatranscriptomic census of active protists in soils. *ISME J.* **9**, 2178–2190 (2015).
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
- Berger, S. A. & Stamatakis, A. Aligning short reads to reference alignments and trees. *Bioinformatics* **27**, 2068–2075 (2011).
- Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**, 538 (2010).
- Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
- Brown, M. W. et al. Phylogenomics places orphan protistan lineages in a novel eukaryotic super-group. *Genome Biol. Evol.* **10**, 427–433 (2018).
- Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
- Crisuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
- Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
- Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).
- Wang, H. C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst. Biol.* **67**, 216–235 (2018).

52. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288 (2009).
53. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).
54. Susko, E. & Roger, A. J. On reduced amino acid alphabets for phylogenetic inference. *Mol. Biol. Evol.* **24**, 2139–2150 (2007).
55. Brown, J. W., Walker, J. F. & Smith, S. A. Phyx: phylogenetic tools for unix. *Bioinformatics* **33**, 1886–1888 (2017).
56. Eddy, S. R. Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
57. Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27**, 221–224 (2010).
58. Foissner, W. & Foissner, I. in *An Illustrated Guide to the Protozoa* 2nd edn (eds Lee, J. J. et al.) 1185–1186 (Society of Protozoologists and Allen Press, Lawrence, 2002).
59. Zolffel, M. & Skibbe, O. Rediscovery of the multiflagellated protist *Paramastix conifera* Skuja 1948 (Protista incertae sedis). *Nova Hedwigia* **65**, 443–452 (1997).



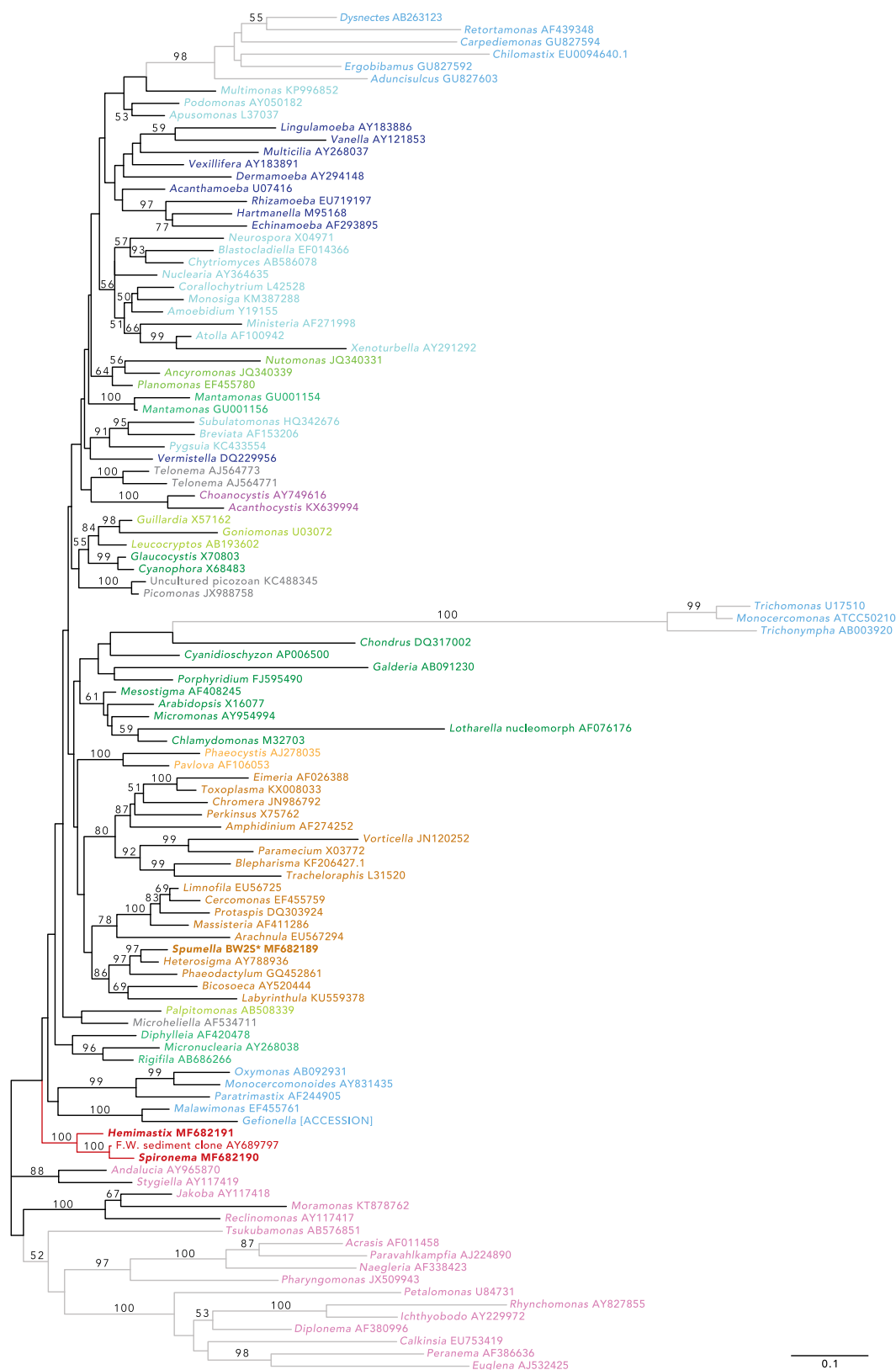
Extended Data Fig. 1 | Light micrographs of studied hemimastigotes. **a–m**, *Spironema* cf. *multiciliatum* (**a**) and *Hemimastix kukwesjijk* (**b–m**) differential interference contrast micrographs of live cells. **a**, Two views of a *Spironema* cf. *multiciliatum* cell, with inset that details the posterior end. Note the nucleus (marked by 'n'), the detail of one of the posterior flagella (marked by an arrow, in the inset) and small contractile vacuole (cv, in inset), as well as posterior tail (line in inset). **b**, **c**, Optical sections through one *H. kukwesjijk* cell, detailing the notches from which flagella emerge (arrowheads), a section through the capitulum (marked with a 'c') and a conspicuous contractile vacuole in the cell posterior (shown in **b**). **d**, Surface view of one of the two thecal plates. **e–g**, Optical cross-sections

of different cells showing the capitulum (**e**), mid-body region with rotationally symmetrical plate overlap (**f**) and the posterior (**g**) with radial arrangement of the posterior-most flagella. **h–j**, Pseudoseries that illustrates the feeding process, showing the progression of prey-ingestion stages. Note the widening capitulum and beginning of formation of the phagocytic vacuole. **k**, Same cell as in **j**, showing the anterior flagella curving forward to surround prey (seen especially in early feeding). **l**, **m**, Dividing cells, showing the diagonal symmetry of short new rows (nr) and longer old rows (or) of flagella, as well as the daughter nuclei (n). Scale bar, 10 μ m.



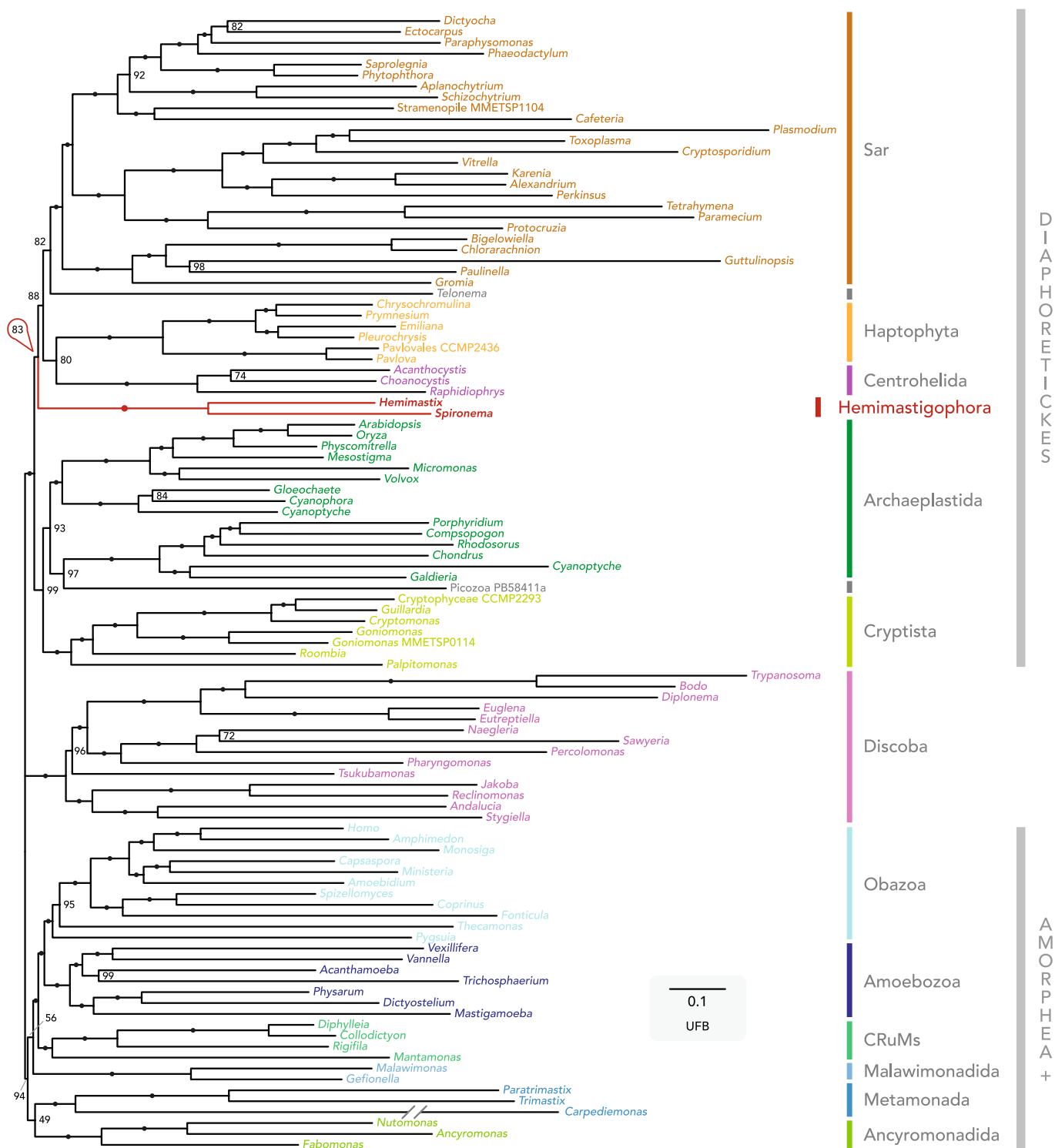
Extended Data Fig. 2 | Scanning electron microscopy images of *H. kukwesjijk*. **a**, Feeding cell, general view (anterior to left; note the prey item attached to capitulum). **b**, Close-up of anterior end showing ingestion in progress at the capitulum. **c**, Discharged extrusomes (ex; triggered

by the fixation process) along margin of the capitulum (compare to undischarged extrusomes in Fig. 1d). **d**, Dividing cells, with the left-most cell clearly showing the old row of full-length flagella (or) and the new row with short flagella (nr). Scale bars, 5 μm (**a**, **d**), 2 μm (**b**, **c**).



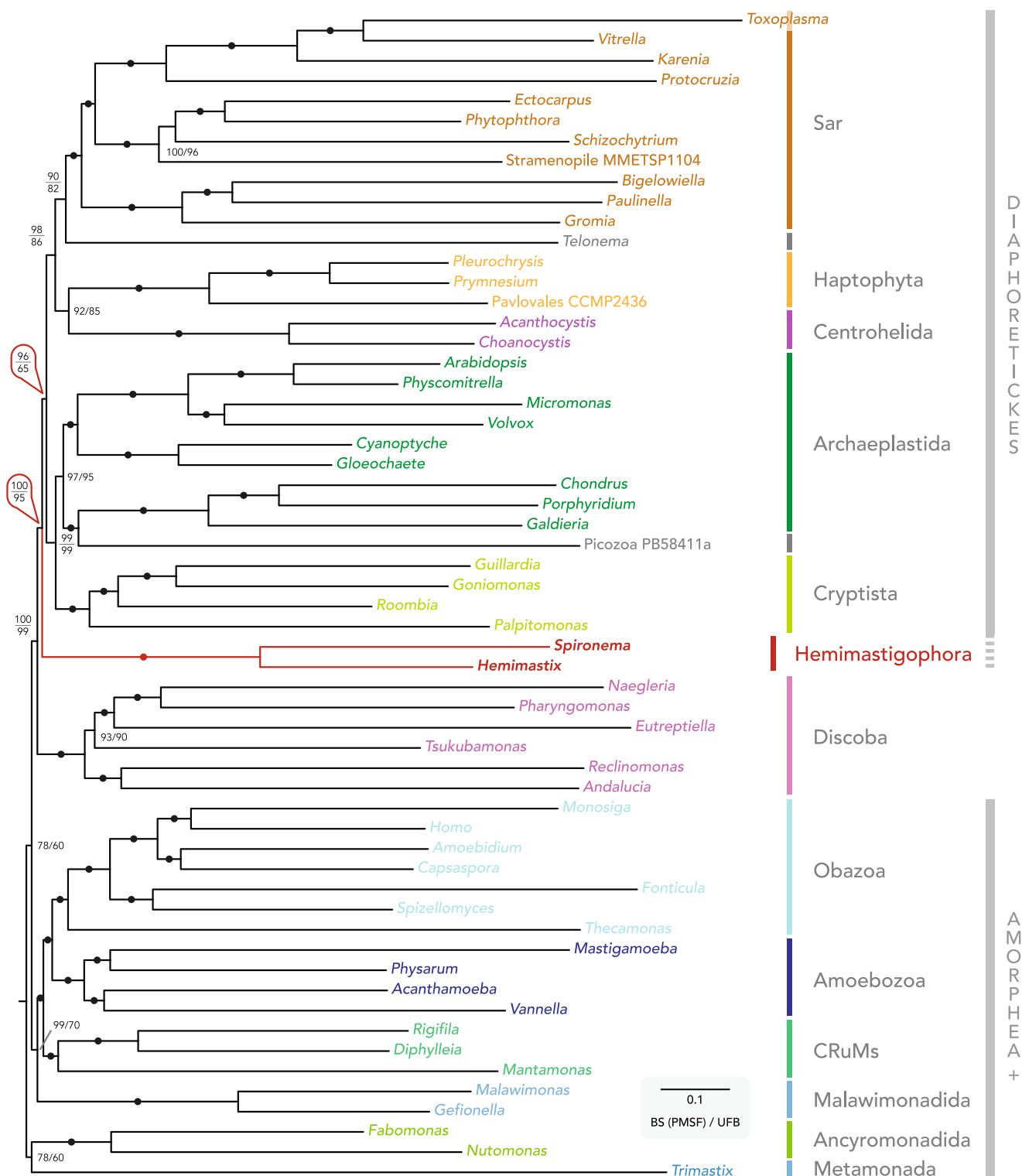
Extended Data Fig. 3 | SSU rDNA phylogeny of eukaryotes. Phylogeny inferred from 111 taxa and 1,252 sites under the GTR + Γ model in RAxML. Hemimastigophora—including *H. kukwesjijk* and *Spironema* cf. *multiciliatum* from this study—are shown in red. Colours of other sequence names correspond to the same taxonomic groupings as in Fig. 3. The sequence of *Spumella* sp. strain BW2S, the prey for *H. kukwesjijk*, is

included and marked with an asterisk. The numbers on branches show bootstrap percentages (1,000 replicates; values below 50% not shown). Branches in grey are half their original length. This tree was the reference phylogeny for pplacer analyses shown in Fig. 2. Scale bar denotes 0.1 expected substitutions per site.



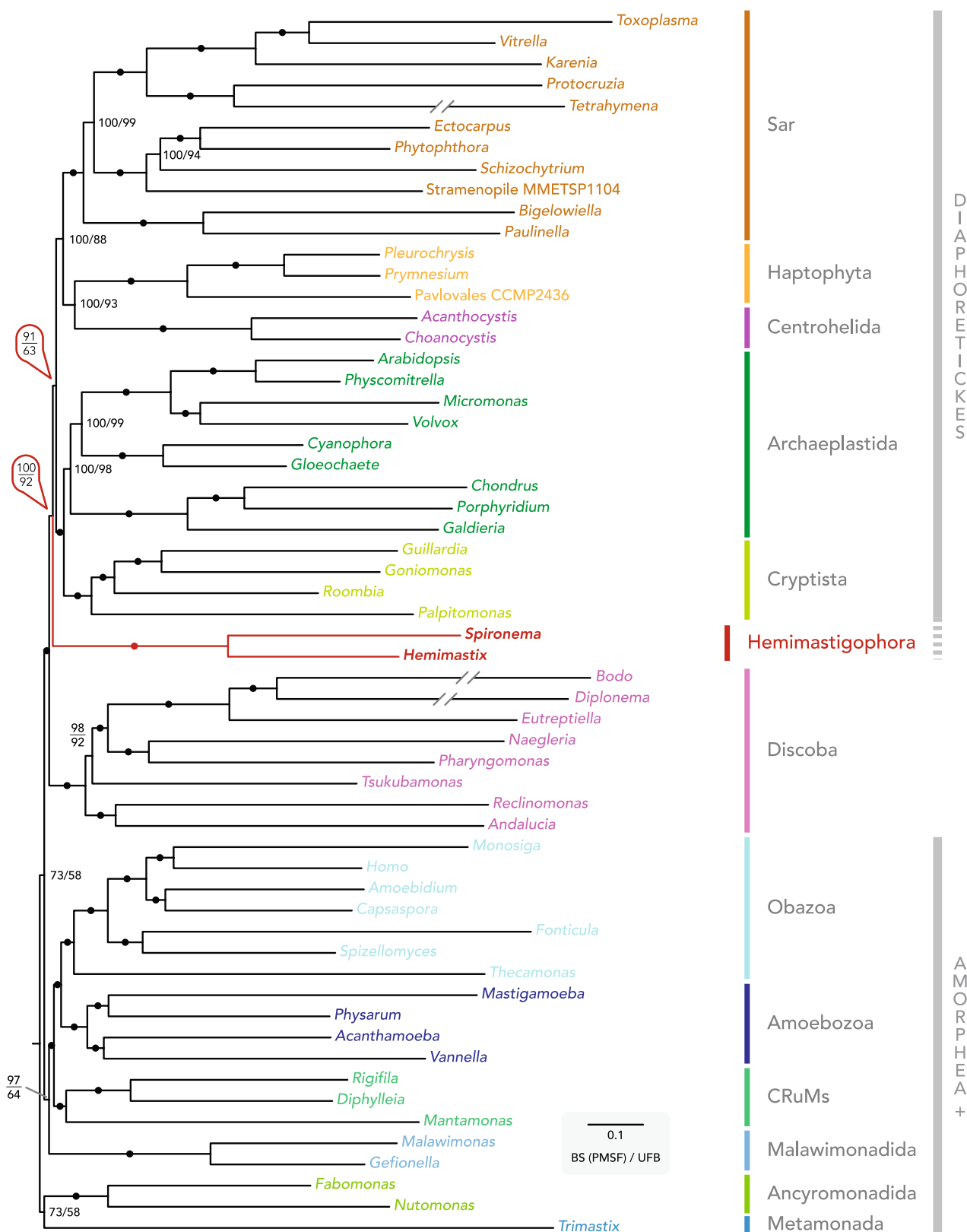
Extended Data Fig. 4 | Unrooted phylogeny of eukaryotes, 104 taxa dataset. Phylogeny inferred from 351 genes, using maximum likelihood under the LG + C60 + F + Γ model. The numbers on branches show ultrafast bootstrap approximation percentages, with filled circles denoting

100% support. The *Carpediemonas* branch is shown reduced by 1/3 of the original length for display purposes. Scale bar denotes 0.1 expected substitutions per site.



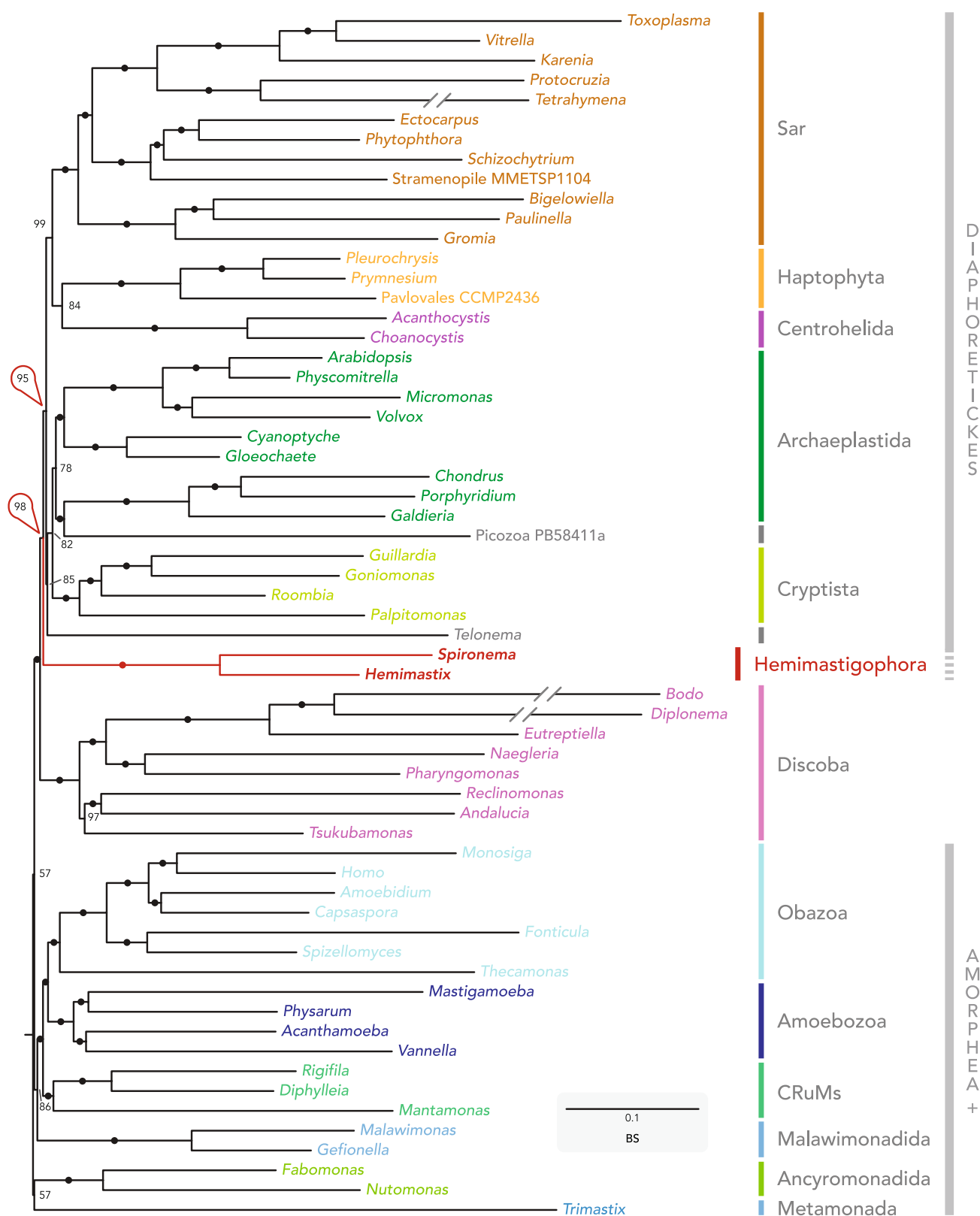
Extended Data Fig. 5 | Unrooted phylogeny using 58-nLB dataset. Phylogeny inferred from 351 genes, using maximum likelihood under the LG + C60 + F + Γ model. The numbers on branches show PMSF bootstrap percentages (bootstrap support PMSF; 200 true bootstrap

replicates), then ultrafast bootstrap approximation percentages (1,000 replicates). Filled circles denote 100% support with both methods. Scale bar denotes 0.1 expected substitutions per site.



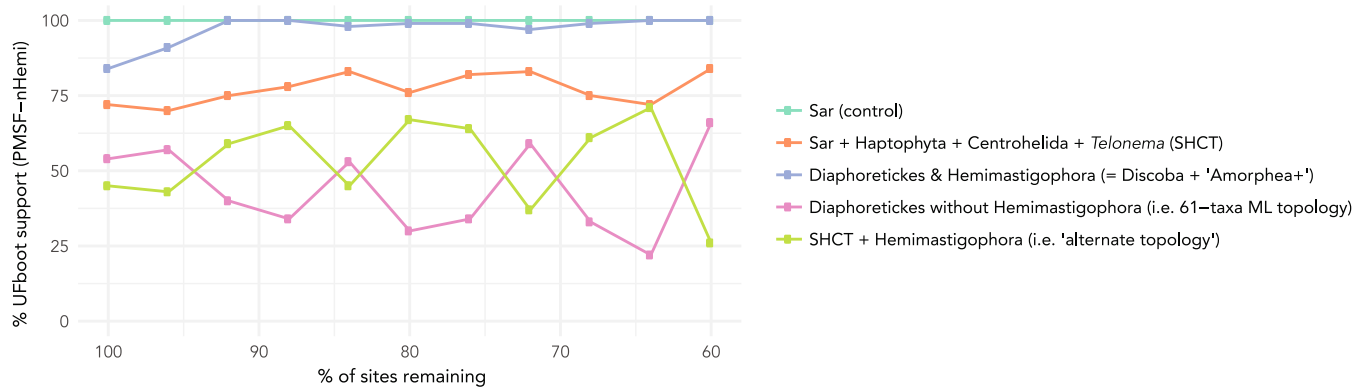
Extended Data Fig. 6 | Unrooted phylogeny using 58-nDP dataset. Phylogeny inferred from 351 genes, using maximum likelihood under the LG + C60 + F + Γ model. The numbers on branches show PMSF bootstrap percentages (bootstrap support PMSF; 100 true bootstrap

replicates), then ultrafast bootstrap approximation percentages (1,000 replicates). Filled circles denote 100% support with both methods. The branches leading to *Bodo*, *Diplonema* and *Tetrahymena* are shown reduced by 1/3. Scale bar denotes 0.1 expected substitutions per site.



Extended Data Fig. 7 | Unrooted phylogeny using 61-SR4 dataset of 61 taxa. Phylogeny inferred from 351 genes, with amino acids recoded as four states, using maximum likelihood under the GTR + R6 + F model. The numbers on branches show bootstrap percentages (500 true bootstrap

replicates). Filled circles represent 100% support. The branches leading to *Bodo*, *Diplonema* and *Tetrahymena* are shown reduced by 1/3. Scale bar denotes 0.1 expected substitutions per site.



Extended Data Fig. 8 | Summary of 61-SFSR analysis. Chart follows the support for several important bipartitions with the sequential removal of the fastest-evolving sites from the 61-taxon, 351-gene dataset. The support values are ultra-fast bootstrap approximation

percentages (1,000 replicates) inferred using maximum likelihood under the LG + C60 + F + Γ -derived PSMF model using a guide tree pruned of hemimastigotes (PMSF-nHEMI, see Methods); these values are not directly comparable to those from the other illustrated analyses.

Transmission of amyloid- β protein pathology from cadaveric pituitary growth hormone

Silvia A. Purro¹, Mark A. Farrow¹, Jacqueline Linehan¹, Tamsin Nazari¹, David X. Thomas¹, Zhicheng Chen², David Mengel², Takashi Saito³, Takaomi Saido³, Peter Rudge¹, Sebastian Brandner^{1,4}, Dominic M. Walsh^{1,2} & John Collinge^{1,*}

We previously reported¹ the presence of amyloid- β protein (A β) deposits in individuals with Creutzfeldt–Jakob disease (CJD) who had been treated during childhood with human cadaveric pituitary-derived growth hormone (c-hGH) contaminated with prions. The marked deposition of parenchymal and vascular A β in these relatively young individuals with treatment-induced (iatrogenic) CJD (iCJD), in contrast to other prion-disease patients and population controls, allied with the ability of Alzheimer's disease brain homogenates to seed A β deposition in laboratory animals, led us to argue that the implicated c-hGH batches might have been contaminated with A β seeds as well as with prions. However, this was necessarily an association, and not an experimental, study in humans and causality could not be concluded. Given the public health importance of our hypothesis, we proceeded to identify and biochemically analyse archived vials of c-hGH. Here we show that certain c-hGH batches to which patients with iCJD and A β pathology were exposed have substantial levels of A β ₄₀, A β ₄₂ and tau proteins, and that this material can seed the formation of A β plaques and cerebral A β –amyloid angiopathy in intracerebrally inoculated mice expressing a mutant, humanized amyloid precursor protein. These results confirm the presence of A β seeds in archived c-hGH vials and are consistent with the hypothesized iatrogenic human transmission of A β pathology. This experimental confirmation has implications for both the prevention and the treatment of Alzheimer's disease, and should prompt a review of the risk of iatrogenic transmission of A β seeds by medical and surgical procedures long recognized to pose a risk of accidental prion transmission^{2,3}.

Human prion diseases occur most commonly as sporadic or inherited conditions but, critically, are also experimentally transmissible, and rare cases are acquired by environmental exposure to infectious prions via diet or medical procedures². This aetiological triad in prion disorders was thought to be unique amongst neurodegenerative diseases, but growing evidence from experimental cellular and animal models has implicated the propagation and spread of multimeric assemblies of misfolded host proteins in the pathogenesis of Alzheimer's, Parkinson's and other neurodegenerative conditions^{2,4}.

Iatrogenic transmission of CJD, an invariably lethal neurodegenerative disease, can result following a range of medical and surgical procedures^{2,5}. The range of incubation periods of acquired prion diseases is known to span more than five decades⁶. Before 1985, when the risk of causing iCJD was not appreciated, children with short stature were treated with growth hormone extracted from large pools of cadaver-derived pituitary glands, some of which would have been infected with prions⁷. More than 200 individuals treated with c-hGH worldwide have died of iCJD. We previously reported moderate to severe grey-matter and vascular A β pathology in four of eight relatively young adults who had died of iCJD following childhood treatment with c-hGH¹. A further two had focal A β pathology and only one was entirely negative for A β . All eight lacked genetic risk factors for Alzheimer's disease

or cerebral A β –amyloid angiopathy (CAA). These findings stood in marked contrast to other prion disease and population controls and suggested that some of the c-hGH with which they were treated was contaminated with A β seeds as well as human prions. While these individuals did not have the full diagnostic neuropathological features of Alzheimer's disease—which also requires the presence of intracellular neurofibrillary tangles—some did have undoubted CAA disease with circumferential vessel-wall degeneration. Had they not died of iCJD at a relatively young age¹, these individuals would have been expected to develop cerebral haemorrhage.

CAA can occur independently of Alzheimer's disease, but at autopsy CAA is detected in the large majority of Alzheimer's cases^{8,9}. That CAA, a pathology that leads to cerebral haemorrhage and dementia, is most often caused by A β deposition in blood vessels is undoubted¹⁰. Indeed, as with Alzheimer's disease, autosomal dominant mutations in, or triplication of, the amyloid precursor protein (APP) gene can cause CAA^{11–13}. The transmissibility of CAA, and potentially Alzheimer's disease, by iatrogenic routes raises important public health issues and would also

Table 1 | Quantification of A β species and tau in c-hGH preparations

| Preparation method | Batch number | A β _{x-40} (pg per vial) | A β _{x-42} (pg per vial) | Tau (pg per vial) |
|--------------------|--------------|---|---|-------------------|
| HWP | 40 | 582 | 116 | 12,411 |
| HWP | 42 | 288 | NQ | 13,631 |
| HWP | 43 | 575 | 112 | 14,581 |
| HWP | 47 | 772 | 108 | 14,569 |
| HWP | 51 | 991 | 136 | 18,155 |
| FL | 4 | NQ | NQ | NQ |
| FL | 5 | NQ | NQ | NQ |
| FL | 6 | NQ | NQ | NQ |
| LJ | 7 | NQ | NQ | NQ |
| LJ | 9 | NQ | NQ | NQ |
| LJ | 10 | NQ | NQ | NQ |
| TPL | 3 | NQ | NQ | NQ |
| TPL | 6 | NQ | NQ | NQ |
| TPL | 14 | NQ | NQ | NQ |
| TPL | 18 | NQ | NQ | NQ |
| TPL | 25 | NQ | NQ | NQ |

With the exception of HWP 42 and HWP 43, all batches were administered to the patients with A β pathology described in ref. ¹; each patient received multiple injections from a number of different batches and preparations, although the table lists only the batches and preparations for which vials were available for us to test. All patients received HWP-prepared c-hGH; only batches 40, 42, 43, 47 and 51 were available. A full list of the number of injections, preparations and batches that each patient received is provided in Extended Data Table 1. 'NQ' indicates that samples did not have quantifiable amounts of analyte. The lowest amount of analyte measurable in a vial is calculated on the basis of the lower limit of quantification for the assay, plus a mathematical adjustment to account for sample dilution. The predicted lowest measurable amounts of A β ₄₀, A β ₄₂ and tau per vial were 148.3 pg, 71.2 pg and 11.4 ng, respectively.

¹MRC Prion Unit at UCL, UCL Institute of Prion Diseases, London, UK. ²Laboratory for Neurodegenerative Research, Ann Romney Center for Neurologic Diseases, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. ³Laboratory for Proteolytic Neuroscience, RIKEN Center for Brain Science, 2-1 Hirosawa, Wako, Japan. ⁴Division of Neuropathology, National Hospital for Neurology and Neurosurgery, London, UK. *e-mail: jc@prion.ucl.ac.uk

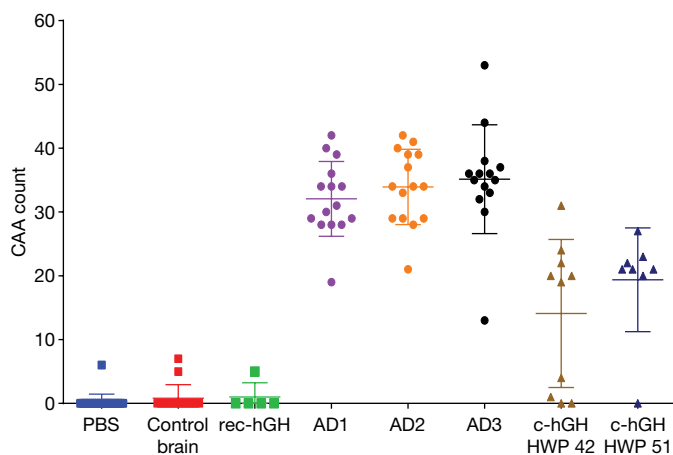


Fig. 1 | Quantification of vessels with CAA in *App*^{NL-F/NL-F} mice following inoculation with Alzheimer's or control human brain, vehicle alone, or recombinant or cadaveric human growth hormone. There were highly significant differences between vehicle (PBS)-inoculated mice and those inoculated with either AD brain homogenates or c-hGH preparations (PBS, $n = 25$; AD1, $n = 15$; PBS versus AD1, $P < 0.0001$; AD2, $n = 15$; PBS versus AD2, $P < 0.0001$; AD3, $n = 14$; PBS versus AD3, $P < 0.0001$; c-hGH HWP 42, $n = 10$; PBS versus HWP 42, $P < 0.0001$; c-hGH HWP 51, $n = 8$; PBS versus HWP 51, $P < 0.0001$; one-way analysis of variance (ANOVA) followed by Dunnett's multiple comparison test). There were no significant differences between PBS-inoculated mice and those inoculated with control human brain homogenate or rec-hGH (PBS, $n = 25$; control brain, $n = 15$; PBS versus control brain, $P = 0.99$; rec-hGH, $n = 5$; PBS versus rec-hGH, $P = 0.99$). Data are expressed as means \pm standard deviation; n = number of mice per group.

indicate a clear shift in understanding their aetiology and suggest new approaches to prevention and treatment^{2,3}. Alternative interpretations of our findings have been proposed^{14,15}, although we have not considered these to be as plausible as the human transmission hypothesis^{16,17}. Given the potential public health importance of our findings, we proceeded to examine experimentally whether c-hGH batches to which these patients were exposed contained viable A β seeding activity, albeit after storage for more than 30 years.

In the UK, 1,883 patients were treated with c-hGH over the period 1958–1985 and 80 have so far developed iCJD (to July 2018), with recent incubation periods exceeding 40 years^{7,18}. During this period of treatment, multiple preparations using several different extraction methods were used, and patients generally received multiple batches from different preparations. However, one preparation, produced by the Hartree-modified Wilhelmi procedure (HWP), was received by all individuals who went on to develop iCJD^{7,18}. It is thought that size-exclusion chromatography, used in non-Wilhelmi preparation methods, may have reduced prion contamination⁷. Fortunately, Public Health England has maintained an archive of vials of c-hGH batches used to treat patients and we were able to obtain vials from a range of batches and production methods to which the eight patients we described¹ were exposed, plus additional vials from two further HWP batches (Table 1 and Extended Data Table 1).

We analysed vial contents biochemically for the presence of A β peptides (x-40 and x-42) and tau protein (Table 1 and Methods). All HWP vials analysed were clearly positive for A β _{x-40} and tau, and all but one were also positive for A β _{x-42} peptides. A β and tau levels in vials from all other c-hGH purification methods examined—FL (Lowry preparation), LJ (Roos method) and TPL (Centre for Applied Microbiological Research)¹⁹—were below the limits of detection. There was therefore unequivocal biochemical evidence for the presence of A β peptides and tau protein in some of the batches (produced by the HWP method) to which iCJD patients with A β pathology were exposed. However, to determine whether seeding activity is present in this material requires a biological rather than a biochemical assay, as the composition and

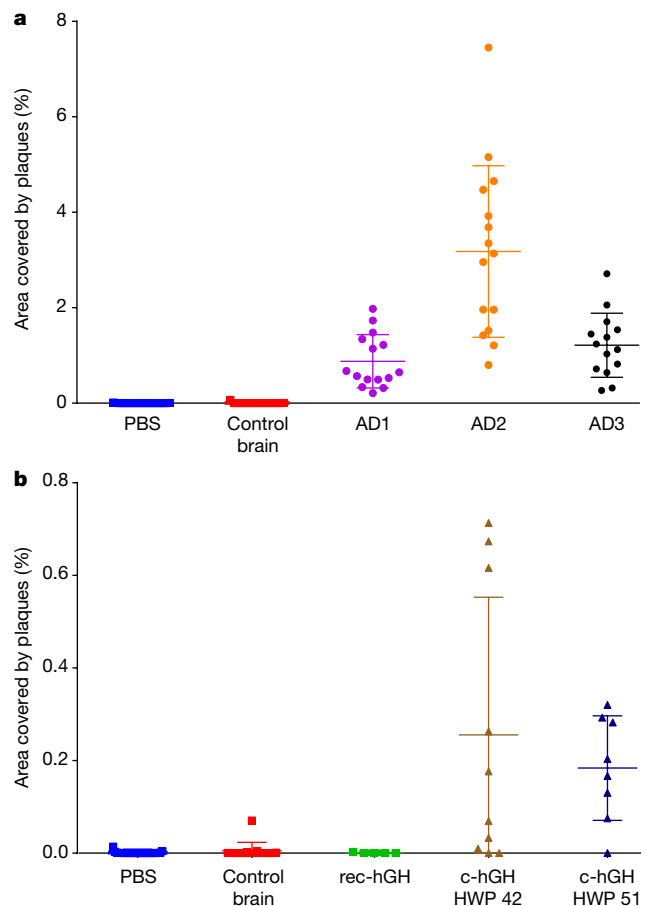


Fig. 2 | Quantification of cerebellar plaque area in *App*^{NL-F/NL-F} mice inoculated with Alzheimer's or control human brain, vehicle alone, or recombinant or cadaveric human growth hormone. The area covered by plaques is expressed as a percentage of the total area: **a**, following inoculation with PBS, control normal brain or AD brain; **b**, following inoculation with PBS, control human brain, rec-hGH or c-hGH. There was no significant difference between PBS-inoculated, control-brain-inoculated or rec-hGH-inoculated mice (PBS, $n = 25$; control brain, $n = 15$; PBS versus control brain, $P = 0.99$; rec-hGH, $n = 5$; PBS versus rec-hGH, $P > 0.99$). However, there were significant differences between PBS-inoculated and AD- or c-hGH-inoculated mice (PBS, $n = 25$; AD1, $n = 15$; PBS versus AD1, $P = 0.007$; AD2, $n = 15$; PBS versus AD2, $P < 0.0001$; AD3, $n = 14$; PBS versus AD3, $P = 0.0002$; c-hGH HWP 42, $n = 10$; PBS versus HWP 42, $P < 0.0001$; c-hGH HWP 51, $n = 8$; PBS versus HWP 51, $P = 0.002$; one-way ANOVA followed by Dunnett's multiple comparison test). Data are expressed as means \pm standard deviation.

structure of seed-competent A β entities is unknown. Indeed, total A β peptide concentrations may be misleading in this regard.

For seeding studies, we used homozygous APP^{NL-F/NL-F} knock-in mice²⁰, which express APP bearing the Swedish (KM670/671NL) and Beyreuther/Iberian (I716F) mutations, with a humanized A β domain; these mice produce the first signs of A β deposition at around six months of age²¹. We conducted extensive in-house time-course studies of uninoculated *App*^{NL-F/NL-F} mice (C57BL/6J background) and confirmed a similar evolution of pathology to that described previously²¹. Inoculating these mice with brain homogenates (1% w/v) prepared from three autopsy-confirmed typical Alzheimer's patients (designated AD 1–3) or a normal control individual, or with vehicle alone (phosphate-buffered saline, PBS)—intracerebrally injected into groups of female *App*^{NL-F/NL-F} mice at 6–8 weeks of age—showed clear seeding of A β pathology from the Alzheimer's cases (Extended Data Fig. 1). Mice were culled at serial time points, namely 2, 7, 15, 30, 45, 60, 90, 120, 240, 360 and 480 days post-inoculation (d.p.i.). Representative

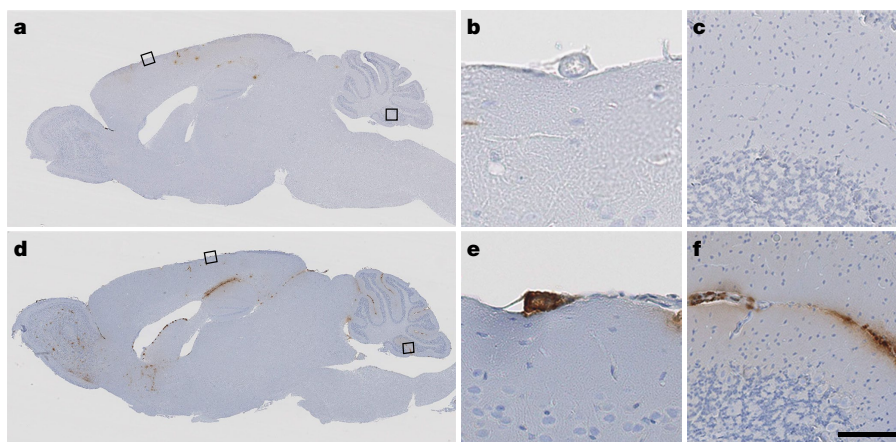


Fig. 3 | A β plaque deposition and CAA in *App*^{NL-F/NL-F} mice following inoculation with AD or control brain. **a–f, *App*^{NL-F/NL-F} mice were inoculated with either human control brain (**a–c**; $n = 15$ mice) or AD brain (**d–f**; $n = 44$ mice) homogenates and culled after 240 days. A β deposition was assessed on sagittal sections (**a**, **d**). CAA (**b**, **e**) and**

cerebellar deposition (**c**, **f**) were evident only in AD-brain-inoculated animals. Boxes denote areas magnified in the middle and right panels. Scale bars represent 1.5 mm for whole sections (**a**, **d**), 25 μ m for CAA (**b**, **e**) and 50 μ m for the cerebellar region (**c**, **f**).

images of A β immunohistochemistry at selected time points are shown in Extended Data Fig. 1. No parenchymal or vascular A β deposits were observed in any mice ($n = 5$ per group) at 2 d.p.i., demonstrating that A β deposition at later time points could not be attributed to persistence of the original inoculum. Meningeal CAA (mainly at the dorsal brain surface) and parenchymal deposition (mainly in the corpus callosum, but also in the cerebellum, hippocampus and cerebral cortex) was detected at 120 d.p.i. in groups inoculated with AD 1–3, but not in those inoculated with PBS or normal brain homogenate ($n = 15$ mice per group; Extended Data Fig. 1). At 240 d.p.i., while PBS- and normal-brain-inoculated mice had almost no amyloid deposits in blood vessels (with minimal deposits seen in 1 out of 25 and 2 out of 15 mice, respectively), AD-brain-inoculated animals had consistent CAA, with ventral meningeal blood vessels (surrounding the olfactory bulb) and many dorsal meningeal vessels affected in all mice ($n = 14$ –15 mice per group; Figs. 1 and 3).

The CAA count was significantly higher in AD-brain-inoculated *App*^{NL-F/NL-F} mice than in PBS-inoculated controls ($P < 0.0001$; Fig. 1). Consistent with previous descriptions of spontaneous pathology in this mouse line²¹, at 240 d.p.i. *App*^{NL-F/NL-F} mice inoculated with PBS or normal brain had only occasional parenchymal plaques in the cerebral cortex and hippocampus, while the cerebellum, olfactory bulb and other areas entirely lacked plaques. However, at this time point widespread parenchymal plaques were evident in the cerebral cortex, hippocampus, corpus callosum, cerebellum and olfactory bulb in all AD-brain-inoculated animals (Fig. 3). The mean percentage area covered by parenchymal plaques was significantly higher in AD-brain-inoculated mice than in PBS-inoculated mice (PBS control versus AD1, $P = 0.017$; AD2, $P < 0.0001$; AD3, $P = 0.0005$; data not shown). The difference in parenchymal A β deposition between AD- and control-inoculated mice was most pronounced in the cerebellum, where deposition was almost completely absent in PBS- or normal-brain-inoculated mice, but marked in AD-inoculated animals (Fig. 2a).

At 360 d.p.i., the localization of plaques observed in AD-brain-inoculated mice was similar but more severe than at 240 d.p.i. Notably, at this time point in PBS- and normal-brain-inoculated mice, CAA was evident only in some dorsal meningeal blood vessels over the cerebral cortex; in marked contrast, in AD-brain-inoculated mice A β deposition was seen in almost all meningeal blood vessels ($n = 15$ mice per group; Extended Data Fig. 1).

To investigate the possible toxicity of intracerebrally administered human growth hormone in mice before we used the scarce c-hGH samples, we inoculated groups of three female C57BL/6J mice with 30 μ l of recombinant human growth hormone (rec-hGH) at 1.2, 3.6 or 11 mg ml⁻¹, corresponding to doses of 0.1, 0.3 and 1 international units

(IU) respectively. There was no evidence of toxicity in any of the mice, which were culled at 240 d.p.i. When mice ($n = 2$) were injected with a higher concentration of recombinant growth hormone (20 mg ml⁻¹; 1.8 IU), they died immediately after the injection.

To establish whether seeding activity was present in A β -positive c-hGH batches, we used vials from c-hGH batches HWP 42 and 51—for which sufficient material was available for inoculation into groups of mice—for similar intracerebral injection into female congenic *App*^{NL-F/NL-F} mice at 6–8 weeks of age. We expected these seeds, if present, to be at a very low titre by comparison with AD-brain homogenate and therefore used the much more efficient transmission route of intracerebral injection of c-hGH, rather than the peripheral injection that patients with iCJD underwent, in order to optimize the chance of detecting seeds in this scarce material²². We also inoculated rec-hGH as a further control, in case growth hormone itself might induce A β deposition. Mice received doses of 0.3 IU of HWP 42, 0.75 IU of HWP 51 or 1 IU of rec-hGH.

As additional experimental controls, AD- and normal-brain 1% w/v homogenate, vehicle alone, HWP 42 and HWP 51 were also injected into wild-type C57BL/6J mice at 6–8 weeks of age. All mice were analysed at 240 d.p.i. As expected, none of the wild-type C57BL/6J mice groups, expressing only murine APP, developed A β deposition ($n = 8$ –10 mice per group; data not shown). Similarly, no cerebellar A β deposits were detected in rec-hGH-inoculated *App*^{NL-F/NL-F} mice ($n = 5$ mice per group) and the CAA score was not statistically different from that of PBS- or normal-brain-inoculated groups (Figs. 1, 2b and 4).

By contrast, CAA and cerebellar A β deposits were clearly evident in *App*^{NL-F/NL-F} mice injected with HWP 42 or HWP 51 (Figs. 1, 2b, 4 and Extended Data Fig. 2), demonstrating the presence of seeding activity in archived HWP c-hGH vials. That the degree of CAA and cerebellar A β deposition in c-hGH-inoculated *App*^{NL-F/NL-F} mice was less pronounced than in those inoculated with AD brain is consistent with the expected much higher seed titre in AD brain than in the archived c-hGH samples. Indeed, it is remarkable that detectable seeding activity has persisted at all after decades of storage.

Our proposal that human transmission of A β pathology had occurred as a result of intramuscular injection of c-hGH is now firmly supported by experimental evidence. While the individuals we described in our earlier report¹ did not meet the full neuropathological criteria for Alzheimer's disease, they might have done so if they had not died of iCJD at a relatively young age. Although tau pathology was not detected in the iCJD patients, it is interesting that the HWP c-hGH batches to which these individuals were exposed also contained biochemically measurable levels of tau. In future studies it will be important to determine whether the tau in c-hGH vials can seed

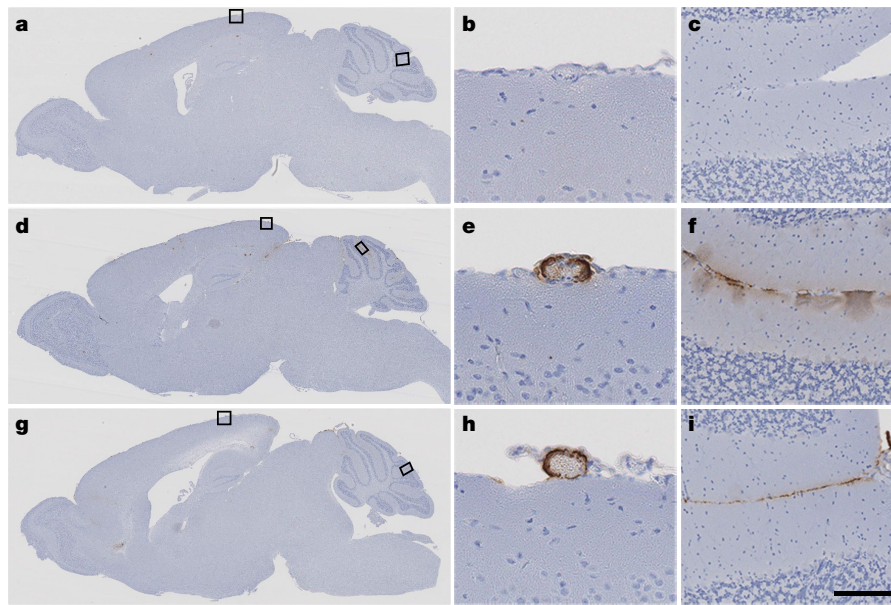


Fig. 4 | A β plaque deposition and CAA in *App*^{NL-F/NL-F} mice following inoculation with cadaveric or recombinant growth hormone preparations. a–i, *App*^{NL-F/NL-F} mice were inoculated with rec-hGH (a–c; *n* = 5 mice), c-hGH batch HWP 42 (d–f; *n* = 10 mice) or c-hGH batch HWP 51 (g–i; *n* = 8 mice) and culled after 240 days. A β deposition

was assessed on sagittal sections (a, d, g). CAA (b, e, h) and cerebellar deposition (c, f, i) were evident in c-hGH- but not rec-hGH-inoculated animals. Boxes denote areas magnified in the middle and right columns. Scale bars represent 1.7 mm for whole sections (a, d, g), 25 μ m for CAA (b, e, h), and 50 μ m for cerebellar regions (c, f, i).

aggregation in mice expressing human tau. However, it is important to emphasize that the seeded A β deposition is not benign: several of these patients had an undoubted disease caused by A β deposition—CAA. This can now be described as iatrogenic CAA (iCAA) and CAA can be considered a transmissible disorder, with attendant public health implications.

After the publication of our original report suggesting human transmission of A β via c-hGH therapy¹—which raised the possibility that it can also be transmitted by other routes known to be a risk for the transmission of CJD prions—there have been several published reports of A β deposition in young individuals following neurosurgical procedures (notably involving dura mater grafting), as well as following c-hGH inoculation^{23–28}. Although we reiterate that there is no suggestion that Alzheimer's disease is contagious, and no supportive evidence from epidemiological studies that it is transmissible (notably by blood transfusion^{29,30}), we consider it important to evaluate the risks of iatrogenic transmission of CAA, and potentially of Alzheimer's disease. Given the lack of disease-modifying therapeutics for Alzheimer's disease and other distressing and fatal neurodegenerative conditions, it will be important to consider introducing improved methods for removing proteopathic seeds from surgical instruments on a precautionary basis.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0790-y>.

Received: 16 August 2018; Accepted: 31 October 2018;

Published online 13 December 2018.

- Jaunmuktane, Z. et al. Evidence for human transmission of amyloid- β pathology and cerebral amyloid angiopathy. *Nature* **525**, 247–250 (2015); erratum **526**, 595 (2015).
- Collinge, J. Mammalian prions and their wider relevance in neurodegenerative diseases. *Nature* **539**, 217–226 (2016).
- Walsh, D. M. & Selkoe, D. J. A critical appraisal of the pathogenic protein spread hypothesis of neurodegeneration. *Nat. Rev. Neurosci.* **17**, 251–260 (2016).
- Jucker, M. & Walker, L. C. Self-propagation of pathogenic protein aggregates in neurodegenerative diseases. *Nature* **501**, 45–51 (2013).
- Brown, P. et al. Iatrogenic Creutzfeldt–Jakob disease at the millennium. *Neurology* **55**, 1075–1081 (2000).

- Collinge, J. et al. Kuru in the 21st century—an acquired human prion disease with very long incubation periods. *Lancet* **367**, 2068–2074 (2006).
- Swerdlow, A. J., Higgins, C. D., Adlard, P., Jones, M. E. & Preece, M. A. Creutzfeldt–Jakob disease in United Kingdom patients treated with human pituitary growth hormone. *Neurology* **61**, 783–791 (2003).
- Charidimou, A. et al. Emerging concepts in sporadic cerebral amyloid. *Brain* **140**, 1829–1850 (2017).
- Biffi, A. & Greenberg, S. M. Cerebral amyloid angiopathy: a systematic review. *J. Clin. Neurol.* **7**, 1–9 (2011).
- Revesz, T. et al. Cerebral amyloid angiopathies: a pathologic, biochemical, and genetic view. *J. Neuropathol. Exp. Neurol.* **62**, 885–898 (2003).
- Levy, E. et al. Mutation of the Alzheimer's disease amyloid gene in hereditary cerebral hemorrhage, Dutch type. *Science* **248**, 1124–1126 (1990).
- Hendriks, L. et al. Presenile dementia and cerebral haemorrhage linked to a mutation at codon 692 of the β -amyloid precursor protein gene. *Nat. Genet.* **1**, 218–221 (1992).
- Selkoe, D. J. & Hardy, J. The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Mol. Med.* **8**, 595–608 (2016).
- Feeney, C. et al. Seeds of neuroendocrine doubt. *Nature* **535**, E1–E2 (2016).
- Adams, H. H. H., A Swanson, S., Hofman, A. & Ikram, M. A. Amyloid- β transmission or unexamined bias? *Nature* **537**, E7–E9 (2016).
- Collinge, J., Jaunmuktane, Z., Mead, S., Rudge, P. & Brandner, S. Collinge et al. reply. *Nature* **537**, E7–E9 (2016).
- Collinge, J., Jaunmuktane, Z., Mead, S., Rudge, P. & Brandner, S. Collinge et al. reply. *Nature* **535**, E2–E3 (2016).
- Rudge, P. et al. Iatrogenic CJD due to pituitary-derived growth hormone with genetically determined incubation times of up to 40 years. *Brain* **138**, 3386–3399 (2015).
- Milner, R. D. Human growth hormone (UK). *Arch. Dis. Child.* **54**, 733–734 (1979).
- Nilsson, P., Saito, T. & Saido, T. C. New mouse model of Alzheimer's. *ACS Chem. Neurosci.* **5**, 499–502 (2014).
- Saito, T. et al. Single App knock-in mouse models of Alzheimer's disease. *Nat. Neurosci.* **17**, 661–663 (2014).
- Eisele, Y. S. et al. Peripherally applied A β -containing inoculates induce cerebral β -amyloidosis. *Science* **330**, 980–982 (2010).
- Frontzek, K., Lutz, M. I., Aguzzi, A., Kovacs, G. G. & Budka, H. Amyloid- β pathology and cerebral amyloid angiopathy are frequent in iatrogenic Creutzfeldt–Jakob disease after dura grafting. *Swiss Med. Wkly* **146**, w14287 (2016).
- Kovacs, G. G. et al. Dura mater is a potential source of A β seeds. *Acta Neuropathol.* **131**, 911–923 (2016).
- Hamaguchi, T. et al. Significant association of cadaveric dura mater grafting with subpial A β deposition and meningeal amyloid angiopathy. *Acta Neuropathol.* **132**, 313–315 (2016).
- Ritchie, D. L. et al. Amyloid- β accumulation in the CNS in human growth hormone recipients in the UK. *Acta Neuropathol.* **134**, 221–240 (2017).
- Duyckaerts, C. et al. Neuropathology of iatrogenic Creutzfeldt–Jakob disease and immunoassay of French cadaver-sourced growth hormone batches suggest possible transmission of tauopathy and long incubation periods for the transmission of Abeta pathology. *Acta Neuropathol.* **135**, 201–212 (2018).

28. Jaunmuktane, Z. et al. Evidence of amyloid- β cerebral amyloid angiopathy transmission through neurosurgery. *Acta Neuropathol.* **135**, 671–679 (2018).
29. Daviglus, M. L. et al. Risk factors and preventive interventions for Alzheimer disease: state of the science. *Arch. Neurol.* **68**, 1185–1190 (2011).
30. O'Meara, E. S. et al. Alzheimer's disease and history of blood transfusion by apolipoprotein-E genotype. *Neuroepidemiology* **16**, 86–93 (1997).

Acknowledgements This work was funded by the UK Medical Research Council (MRC); the National Institute of Health Research (NIHR) University College London Hospitals (UCLH)/University College London (UCL) Biomedical Research Centre; the Leonard Wolfson Experimental Neurology Centre; and a grant to D.M.W. from the National Institute on Aging (AG046275). We thank the Queen Square Brain Bank for Neurological Disorders (supported by the Reta Lila Weston Trust for Medical Research, the Progressive Supranuclear Palsy (Europe) Association and the MRC) at the UCL Institute of Neurology, University College London; and the Oxford Brain Bank (supported by the MRC, the NIHR Oxford Biomedical Research Centre and the Brains for Dementia Research programme, jointly funded by Alzheimer's Research UK and Alzheimer's Society) for providing the UK human brain tissue samples. We thank M. Ellis for image analysis; Z. Jaunmuktane for advice on CAA scoring; and G. Graham, C. Fitzhugh, R. Labesse-Garbal and other staff of the MRC Prion Unit Biological Services facility for animal inoculation, observation and care. We thank M. Farmer and E. Quarterman for technical assistance; O. Avwenagha and J. Wadsworth for assistance in selecting and processing tissue samples; and E. Noble for assistance with assay development. We thank

P. Adlard for help in identifying growth-hormone batches for this study and M. Sutton for providing c-hGH vials from archived stores at Public Health England Porton Down. Antibodies m266, 2G3 and 21F12 were gifts from P. Seubert and D. Schenk, Elan Pharmaceuticals.

Author contributions S.A.P. and M.A.F. coordinated animal experiments and performed data analysis. J.L., T.N. and S.B. performed neuropathological analysis. D.X.T., Z.C., D.M. and D.M.W. performed and analysed biochemical assays. T.Saito and T.Saido provided NL-F mice. P.R. coordinated the identification and sourcing of relevant archival c-hGH batches. J.C. oversaw the study and drafted the manuscript with contributions from all authors.

Competing interests J.C. is a shareholder and director of D-Gen Limited, an academic spin-out company working in the field of prion-disease diagnosis, decontamination and therapeutics.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0790-y>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0790-y>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to J.C.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Use of human tissues and research ethics. This study was carried out following ethics approval from the North East–Newcastle and North Tyneside 2 Research Ethics Committee (REC), reference 11/NE/0348, and the London Queen Square REC, reference 03/N038. The storage and biochemical analysis of human tissue samples and the transmission studies involving mice were performed in accordance with informed consent from all patients, from a person in a qualifying relationship to the deceased, or from a legal representative, in accordance with applicable UK legislation and regulatory codes of practice.

Anonymized post-mortem brain samples (three neuropathologically confirmed cases of Alzheimer's disease and one control with no signs of neurodegenerative disease) were provided under a material transfer agreement from the Queen Square Brain Bank for Neurological Disorders, UCL Institute of Neurology and Oxford Brain Bank, Oxford University Hospitals NHS Trust. Samples were obtained and used in accordance with the requirements of each providing tissue bank.

Sourcing of archived c-hGH material. Human cadaveric pituitary-derived growth hormone material, from batches manufactured in the mid-1980s, is stored at Public Health England (PHE) under contract from the Department of Health and Social Care, with accompanying batch manufacturing records where available. Material from specific manufactured batches was supplied from this archive for the purposes of our study, on the request of the MRC Prion Unit at UCL and with the approval of the Department of Health and Social Care. All material has been stored at ambient temperature in sealed vials since the date of transfer to PHE.

Biochemical analysis of c-hGH vials. The contents of each c-hGH vial were resuspended directly in 316 μ l 6 M guanidine hydrochloride and all analyses were conducted with the investigator blinded to sample identity. For determination of $A\beta_{x-40}$ concentrations, samples were diluted 12-fold before analysis on a Meso Scale Diagnostics platform, using anti- $A\beta$ antibody 266 for capture and biotinylated 2G3 antibody for detection³¹. $A\beta_{x-42}$ concentrations were determined following 72-fold dilution on an Erenna instrument (Quanterix, Lexington, MA, USA) using anti- $A\beta$ antibody 266 for capture and fluorescently labelled 21F12 for detection³¹. Tau levels were determined by enzyme-linked immunosorbent assay (ELISA) following 144-fold dilution, with anti-tau antibody BT2 (Thermo Fisher Scientific) used for capture and alkaline-phosphatase-conjugated Tau5 antibody (Thermo Fisher Scientific) for detection³². We defined lower limits of quantification (LLOQs) for the determination of $A\beta_{x-40}$ and tau levels as the lowest standards with a signal higher than the average signal for the blank plus 9 standard deviations (s.d.), allowing a percentage recovery of $100 \pm 20\%$ or more, and a coefficient of variance (CV) of 20% or less. The LLOQ for $A\beta_{x-42}$ was defined as the lowest interpolated standard that provided a signal twofold that of the background with a percentage CV of 20% or less, and allows a percentage recovery of $100 \pm 20\%$ or more.

Mouse transmission studies. Mouse studies were performed under approval and licence granted by the UK Home Office (Animals (Scientific Procedures) Act 1986), project licence number 70/9022, and conformed to UCL institutional and Animal Research: Reporting of In Vivo Experiments (ARRIVE) guidelines (<http://www.nc3rs.org.uk/arrive-guidelines>).

We used homozygous APP NL-F knock-in mice²⁰ that express APP bearing the Swedish (KM670/671NL) and Beyreuther/Iberian (I716F) mutations, and in which the $A\beta$ domain has been humanized. These were speed-backcrossed (Jackson Laboratories, USA) and maintained on an inbred C57BL/6J background, and used as homozygotes (designated *App*^{NL-F/NL-F}). Wild-type C57BL/6J mice were purchased from Jackson Laboratories via Charles River Laboratories. Mouse genotype was confirmed by polymerase chain reaction (PCR) of ear-punch DNA, and mice were uniquely identified by subcutaneous transponders. All of the mice used were female, as is our standard practice for long-term prion-transmission experiments, for reasons of consistency, animal welfare and logistics: increased fighting amongst groups of males requires them to be housed separately.

Mice (female, aged 6–8 weeks) were randomly assigned to experimental groups, anaesthetized with a mixture of halothane and O_2 , and intracerebrally inoculated into the right hemisphere in the parietal region with 30 μ l of a 1% (w/v) human brain homogenate prepared in Dulbecco's PBS lacking Ca^{2+} or Mg^{2+} ions (D-PBS), vehicle (D-PBS) alone, 11 mg ml⁻¹ rec-hGH (Humatrope Eli Lilly, UK), or c-hGH material prepared in D-PBS. For preparation of human brain homogenate, grey matter was dissected from frontal cortex samples of one healthy control brain and three Alzheimer's cases (AD 1–3), homogenized using glass grinders with D-PBS at 10% (w/v), and subsequently diluted at 1% in D-PBS to inoculate the mice. Levels of $A\beta_{40}$ and $A\beta_{42}$ in the 10% homogenates were quantified using a V-Plex $A\beta$ Peptide Panel 1 6E10 kit (Meso Scale Diagnostics platform), using anti- $A\beta_{40}$ and anti- $A\beta_{42}$ monoclonal antibodies for capture and anti- $A\beta$ antibody 6E10 for detection. The healthy control brain had 2 ± 0.6 ng ml⁻¹ $A\beta_{40}$ and 6.7 ± 2.8 ng ml⁻¹ $A\beta_{42}$. AD1 had 13.7 ± 0.6 ng ml⁻¹ $A\beta_{40}$ and 30.5 ± 1.4 ng ml⁻¹ $A\beta_{42}$. AD2 had 160.1 ± 10.7 ng ml⁻¹ $A\beta_{40}$ and 43.4 ± 10.6 ng ml⁻¹ $A\beta_{42}$. Finally, AD3 had 14.2 ± 0.3 ng ml⁻¹ $A\beta_{40}$ and 57.7 ± 2.1 ng ml⁻¹ $A\beta_{42}$

(mean \pm s.d.). Note that the total $A\beta$ peptide concentrations determined by biochemical assay may not relate to $A\beta$ seeding activity.

For transmission studies, each c-hGH vial was resuspended in 200 μ l D-PBS and the contents of six vials from each batch were pooled before inoculation. This corresponds to each mouse receiving 0.75 IU of HWP51, 0.3 IU of HWP42 or 1 IU of rec-hGH. HWP42 was labelled as containing 2 IU per vial, whereas the HWP51 was labelled as containing 5 IU per vial; hence, although we used the same number of vials from each batch for inoculation, the dose of growth hormone was different. Calculated amounts of $A\beta$ peptides in each 30- μ l inoculum were as follows: HWP 42— $A\beta_{x-40}$, 43 pg; $A\beta_{x-42}$, below limit of quantitation; HWP 51— $A\beta_{x-40}$, 149 pg; $A\beta_{x-42}$, 20 pg. Inocula were prepared following strict biosafety protocols in a microbiological containment level III laboratory, and inoculations performed within a class I microbiological safety cabinet, using disposable equipment to prepare each inoculum. Safety cabinets were decontaminated before preparing inocula to avoid cross-contamination. Mice were culled at 8 months post-inoculation by exposure to CO_2 ; brains were then removed and prepared for immunohistochemistry.

Antibodies and immunohistochemistry. Tissue was fixed in 10% buffered formal saline and incubated in 98% formic acid for 1 h. Following further washing in 10% buffered formal saline, tissue samples were processed and paraffin wax embedded. Serial sections of 5 μ m nominal thickness were taken. $A\beta$ deposition was visualized using biotinylated 82E1 (catalogue number 10326, IBL, Japan) as the primary antibody, using a Ventana Discovery automated immunohistochemical staining machine (Roche, Burgess Hill, UK) and proprietary solutions. Visualization was accomplished with development of 3',3'-diaminobenzidine tetrahydrochloride as the chromogen (DAB Map Kit, Ventana Medical Systems). Haematoxylin was used as the counterstain.

Image capture. Histological slides were digitized on a LEICA SCN400F scanner (LEICA Milton Keynes, UK) at $\times 40$ magnification and 65% image-compression setting during export. Slides were archived and managed on LEICA Slidepath (LEICA Milton Keynes, UK).

Quantification of CAA and parenchymal $A\beta$ protein. All immunohistochemical quantification was performed blind to experimental group. CAA was present in small meningeal vessels and occasionally in small superficial cortical vessels. Because of the small size of meningeal vessels, reliable automated image analysis of CAA was not possible, and negative and positive vessels were quantified by visual inspection. One paramedian (approximately 200 μ m) sagittal section per mouse was analysed. The extent of the CAA was determined by counting the number of $A\beta$ -negative and -positive blood vessels in six anatomical areas of the meninges covering the dorsal part of the brain, including the olfactory bulb and the cerebellum.

Parenchymal $A\beta$, present in the form of diffuse deposits or as plaques, is amenable to image quantification on whole-slide images as described previously¹. All $A\beta$ -immunostained sagittal sections (approximately 200 μ m parasagittal) of whole mouse brains were digitized as described above. Digital image analysis on whole slides was performed using Definiens Developer XD 2.6 (Definiens, Munich, Germany). Initial tissue identification was performed using a resolution equivalent to $\times 10$ magnification, and stain detection was performed at $\times 20$ magnification. Regions of interest (ROIs) were manually selected to separate the cortex, hippocampus and cerebellum; larger artefacts were also manually selected for exclusion from analysis. Tissue detection and initial segmentation was done to identify all tissue within the image, separating the sample from background and non-tissue regions for further analysis. This separation was based on identification of the highly homogeneous relatively bright/white region of background present at the perimeter of each image. A composite raster image produced by selecting the lowest pixel value from the three constituent colour layers (RGB colour model) provided a greyscale representation of brightness. The mean brightness of this background region was used to exclude all background regions from further analysis.

Stain detection (brown) was based on transformation of the RGB colour model to a hue-saturation-density (HSD) representation³³. This provides a raster image of the intensity of each colour of interest (brown and blue). A number of fixed thresholds (T_x) was then used to identify areas of interest (A_x). The thresholds used were in arbitrary units (AU), with a scale of 0 AU to 3 AU in HSD images. The threshold $T_{\text{brown stain}}$ was allocated the value 0.15 AU, with all pixels above this threshold classed as 'stain' (A_{stain}) and those below as 'unstained' ($A_{\text{unstained}}$). A_{stain} was excluded if the intensity of blue staining was not significantly lower than the intensity of brown stain (that is, if the difference was less than 0.1 AU), in order to remove generically dark areas. The remaining A_{stain} areas were further categorized using a threshold $T_{\text{dark brown}}$ of 0.5 AU, to give A_{light} and A_{dark} .

Plaques were then constructed from these A_{light} and A_{dark} objects. Each A_{dark} area was classified as a plaque seed; these were then grown into all surrounding A_{light} areas to give $A_{\text{potential plaque}}$ (constructed of A_{light} and A_{dark}) and $A_{\text{non-plaque}}$ (constructed of only A_{light}). Several exclusions were then applied, as follows. Any $A_{\text{potential plaque}}$ regions with an area less than 20 μm^2 (see below) or containing fewer than three pixels previously classified as A_{dark} were reclassified as $A_{\text{non-plaque}}$.

$A_{\text{potential plaque}}$ areas with a relatively high stain intensity (mean brown (\bar{B}) intensity higher than 0.35 AU) and low variation in brown stain level (standard deviation in brown stain ($B\delta$) below 0.1 AU) were removed into $A_{\text{unstained}}$ as artefacts. $A_{\text{potential plaque}}$ regions with a \bar{B} higher than 0.5 AU and $B\delta$ below 0.25 AU were reclassified as $A_{\text{non-plaque}}$, as they displayed an uncharacteristically dark and varied stain character for plaques. $A_{\text{potential plaque}}$ regions with an area of less than $40 \mu\text{m}^2$ and elliptic character of less than 0.2 (scale of 0 (random shape) to 1 (perfect circle)) were reclassified as $A_{\text{non-plaque}}$. $A_{\text{potential plaque}}$ regions with an area greater than $40 \mu\text{m}^2$ and relative proportion of A_{dark} greater than 70% were reclassified as $A_{\text{non-plaque}}$. The remaining $A_{\text{potential plaque}}$ regions give our final A_{plaque} .

For each ROI, the total area analysed ($A_{\text{unstained}} + A_{\text{plaque}} + A_{\text{non-plaque}}$), A_{plaque} , $A_{\text{non-plaque}}$ and number of A_{plaque} were exported and the percentage area covered by A_{plaque} was determined.

To establish the optimal minimum area for $A_{\text{potential plaque}}$, analyses using 10, 20, 30, 40 and $50 \mu\text{m}^2$ were performed, with a minimum area of $20 \mu\text{m}^2$ being selected. **Statistical analysis and reproducibility.** All statistical analysis and graphs were generated using the package GraphPad PRISM v6 (GraphPad Software, La Jolla, USA). Error bars on graphs denote standard deviations, with statistical significance determined by one-way ANOVA followed by Dunnett's multiple comparison test

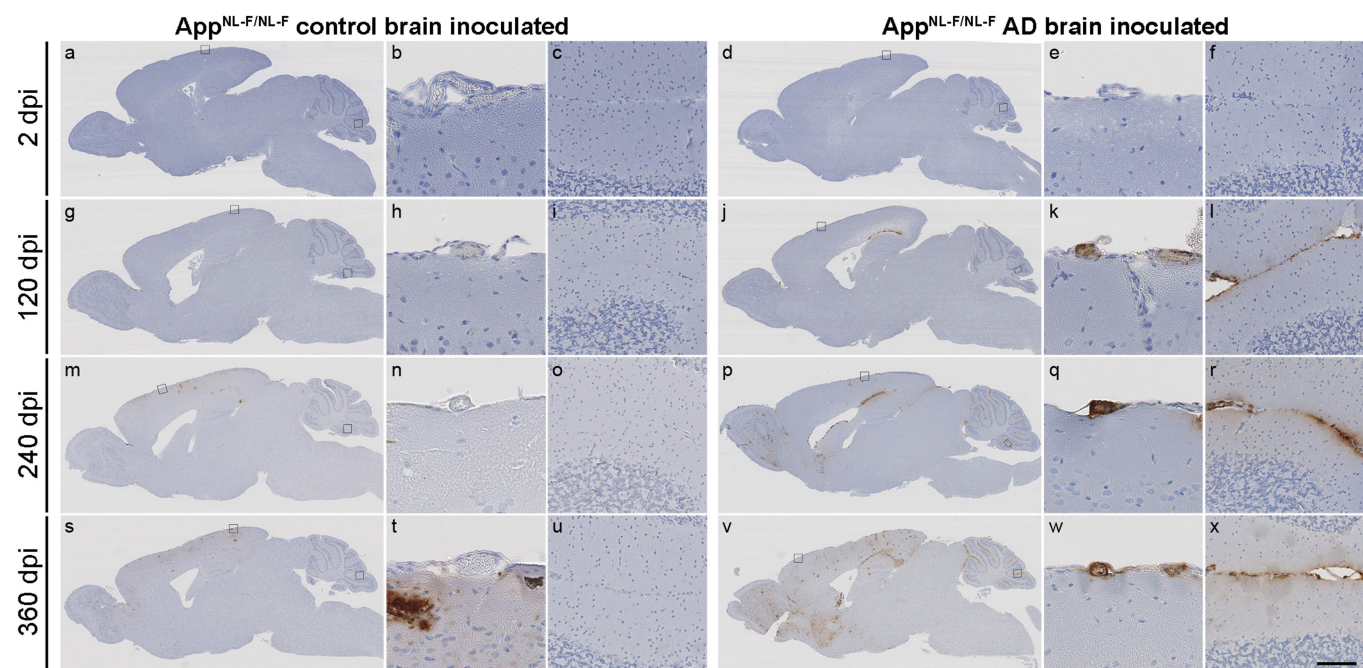
(two-tailed). Statistical significance was set to $P < 0.05$. Experiments with mice were performed only once to avoid unnecessary use of animals, and biochemical assays were not replicated because of the scarcity of cadaveric human growth hormone.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability statement

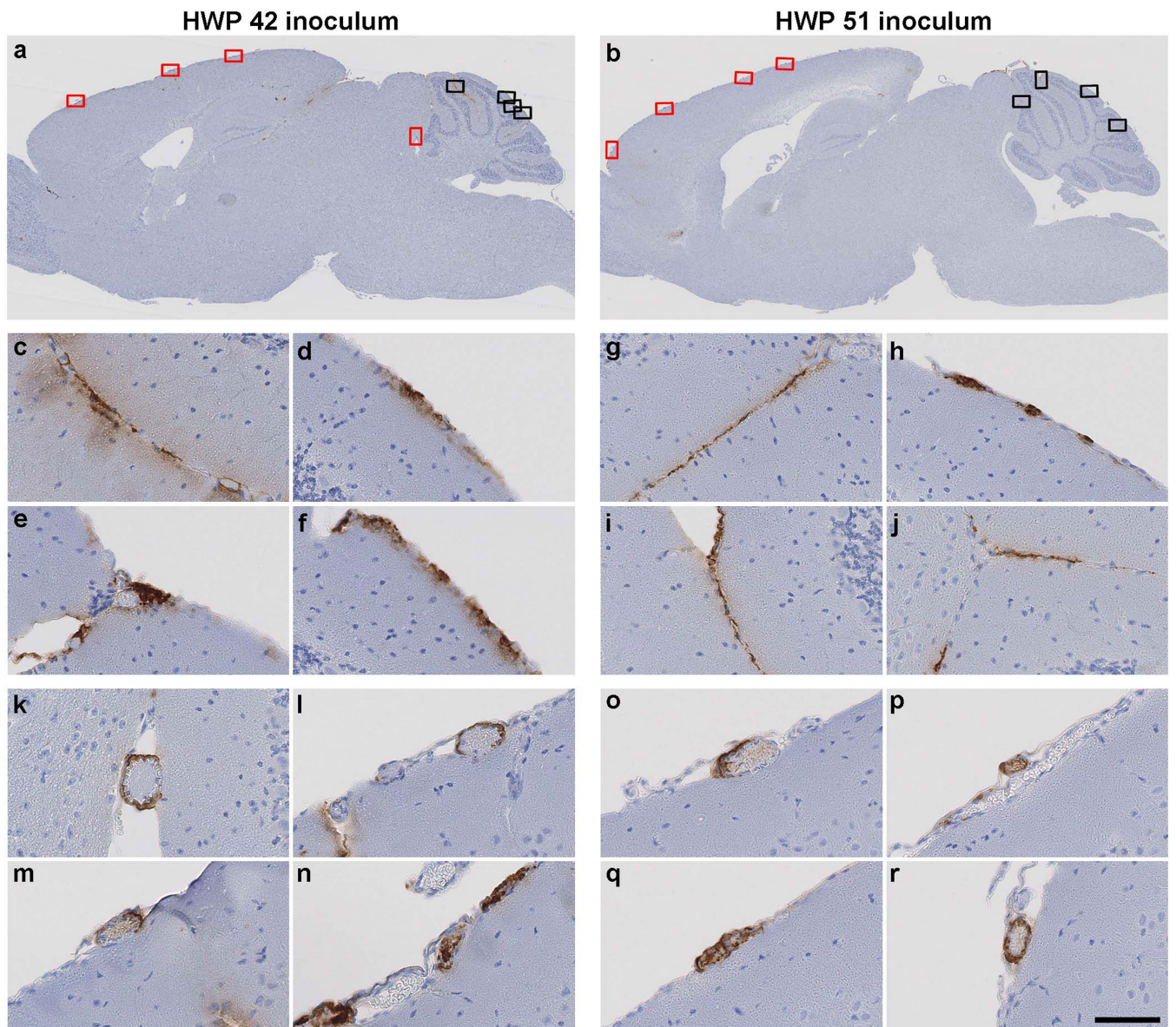
Source data for Figs. 1 and 2 are available from the corresponding author upon reasonable request.

31. Mably, A. J. et al. Anti-A β antibodies incapable of reducing cerebral A β oligomers fail to attenuate spatial reference memory deficits in J20 mice. *Neurobiol. Dis.* **82**, 372–384 (2015).
32. Kanmert, D. et al. C-terminally truncated forms of tau, but not full-length tau or its C-terminal fragments, are released from neurons independently of cell death. *J. Neurosci.* **35**, 10851–10865 (2015).
33. van der Laak, J. A., Pahlplatz, M. M., Hanselaar, A. G. & de Wilde, P. C. Hue-saturation-density (HSD) model for stain recognition in digital images from transmitted light microscopy. *Cytometry* **39**, 275–284 (2000).



Extended Data Fig. 1 | Time course of CAA and A β deposition in control- and AD-brain-inoculated *App*^{NL-F/NL-F} mice. Mice were inoculated with either control-brain homogenates (a–c, g–i, m–o, s–u) or AD-brain homogenates (d–f, j–l, p–r, v–x) and culled at the stated times. A β deposition was assessed on sagittal sections (a, d, g, j, m, p, s, v).

CAA (b, e, h, k, n, q, t, w) and cerebellar deposition (c, f, i, l, o, r, u, x) were evident only in AD-brain-inoculated animals. Boxes denote areas magnified to the right. Scale bars represent 1.4 mm for whole sections (a, d, g, j, m, p, s, v), 25 μ m for CAA (b, e, h, k, n, q, t, w), and 50 μ m for the cerebellar region (c, f, i, l, o, r, u, x).



Extended Data Fig. 2 | Aβ plaques and CAA in *App*^{NL-F/NL-F} mice following inoculation with c-hGH preparations. *App*^{NL-F/NL-F} mice were inoculated with c-hGH batch HWP 42 (a, c–f, k–n) or HWP 51 (b, g–j, o–r) and culled after 240 days. Aβ deposition was assessed on sagittal

sections (a, b). Black and red boxes denote areas magnified to better show cerebellar Aβ deposits (c–j) and CAA (k–r), respectively, in the middle and lower panels. Scale bars represent 1.1 mm for whole sections (a, b) and 50 μm for the cerebellar region and CAA (c–r).

Extended Table 1 | c-hGH preparations and batches received by each patient

| | Cadaveric Human Growth Hormone Preparations | | | | | | |
|------------------------|--|---------|---|----------------|-----------------------------------|-------------------------|----------------------------|
| <i>Patient number*</i> | HWP | K | FL | LJ | TPL | R | |
| 1 | HWP 00, 44, 45, 51 | K 79972 | FL 6 | LJ 4, 7, 9, 10 | TPL 18 | | 1 batch hGH unspecified |
| 2 | HWP 13 batches unspecified | | FL 1, 4, 8, 10 | | TPL 4 and 2 batches unspecified | R 2 batches unspecified | |
| 3 | HWP 9, 10, 15, 19, 20, 28, 29, 31, 40 | K 79250 | FL 4, 8, 9 | LJ 5, 6, 8 | TPL 14, 15 1 batch unspecified | R 15, 16 | |
| 4 | HWP 9, 10, 15, 44, 51 | K 79972 | FL 5, 6, 7 and 4 batches unspecified | LJ 4, 7, 8, 10 | TPL 3, 6, 12, 18, 21, 22, 25 | R 15, 16 | 10 batches hGH unspecified |
| 5 | HWP 11, 21, 23, 28, 29, 38, 40, 51 and 1 batch unspecified | K 79250 | FL 1, 4, 8, 9, 10 and 1 batch unspecified | LJ 4, 5 | TPL 6 | R 18, 19 | |
| 6 | HWP 00, 44, 51 and 4 batches unspecified | K 79972 | | LJ 3, 7 | | | 1 batch hGH unspecified |
| 7 | HWP 00, 45, 47 and 1 batch unspecified | | FL 5, 6 | | | | |
| 8 | HWP 00, 38, 44, 45 | | FL 5, 6 | | | | 2 batches hGH unspecified |

Patient number refers to the patients described in ref. ¹.

c-hGH preparations were as follows: HWP, Hartree-modified Wilhelmi preparation; K, Kabi commercial preparation; FL, St Bartholomew's Hospital preparation using Roos–Lowry method; LJ, commercial preparation using Roos method; TPL, Centre for Applied Microbiology and Research, Porton Down preparation; R, Raben preparation. The final column shows where c-hGH was given but the type of preparation was not specified on medical records.

Distinct activity-gated pathways mediate attraction and aversion to CO₂ in *Drosophila*

Floris van Breugel^{1,2}, Ainul Huda¹ & Michael H. Dickinson^{1*}

Carbon dioxide is produced by many organic processes and is a convenient volatile cue for insects¹ that are searching for blood hosts², flowers³, communal nests⁴, fruit⁵ and wildfires⁶. Although *Drosophila melanogaster* feed on yeast that produce CO₂ and ethanol during fermentation, laboratory experiments^{7–12} suggest that walking flies avoid CO₂. Here we resolve this paradox by showing that both flying and walking *Drosophila* find CO₂ attractive, but only when they are in an active state associated with foraging. Their aversion to CO₂ at low-activity levels may be an adaptation to avoid parasites that seek CO₂, or to avoid succumbing to respiratory acidosis in the presence of high concentrations of CO₂ that exist in nature^{13,14}. In contrast to CO₂, flies are attracted to ethanol in all behavioural states, and invest twice the time searching near ethanol compared to CO₂. These behavioural differences reflect the fact that ethanol is a unique signature of yeast fermentation, whereas CO₂ is generated by many natural processes. Using genetic tools, we determined that the evolutionarily conserved ionotropic co-receptor IR25a is required for CO₂ attraction, and that the receptors necessary for CO₂ avoidance are not involved in this attraction. Our

study lays the foundation for future research to determine the neural circuits that underlie both state- and odorant-dependent decision-making in *Drosophila*.

D. melanogaster feed, mate and deposit eggs on rotting fruit. Between 10 and 14 days later, the next generation of flies must locate a fresh ferment. Because of the high volatility of CO₂, the emission of CO₂ is greatest near the start of fermentation⁸, whereas ethanol emission increases more slowly (Extended Data Fig. 1a). Other odours associated with fermentation (for example, acetic acid and ethyl acetate) form later, when bacteria break down ethanol. In trap assays, *Drosophila* show a preference for two-day-old apple juice ferments compared to older solutions (Extended Data Fig. 1b, c), which suggests that they might be attracted to CO₂. Although it is difficult to estimate concentrations of CO₂ in wild ferments, we measured the CO₂ concentration in bottles commonly used to rear flies to be 0.5–1% (Extended Data Fig. 1d–g).

This evidence that CO₂ might attract *Drosophila* contradicts previous studies conducted using small chambers^{7–12}. To study how flies respond to odours under more-ethological conditions, we recorded

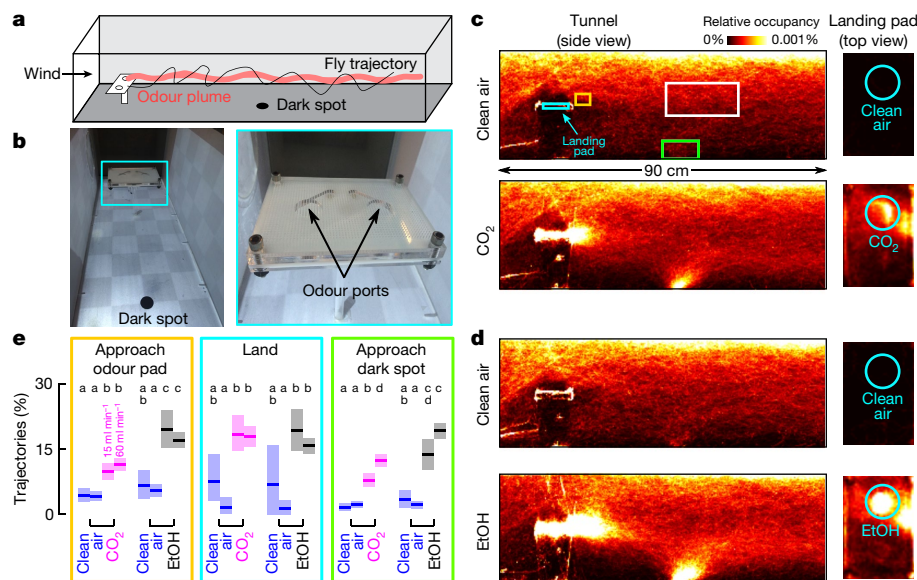


Fig. 1 | *Drosophila* are attracted to ethanol and CO₂ in flight. **a**, Diagram of wind tunnel. **b**, Photograph of the wind tunnel and odour-emitting landing platform. **c**, **d**, Heat maps indicating relative occupancy of flies in the presence of either CO₂ or ethanol. Cohorts of 12 flies were introduced into the wind tunnel and their behaviour recorded over 16 h. Throughout the experiment, 100 ml min⁻¹ of clean air emerged from both odour ports. For 30 min every hour, 60 ml min⁻¹ of either CO₂ or clean air bubbled through 100% ethanol was added to one odour port. Control data come from segments with clean air. Number of cohorts: 9 (CO₂), 6 (ethanol). Number of trajectories: 59,970–101,539 per panel. **e**, Percentage of

trajectories that passed through one of the coloured volumes shown in **c** (gold, cyan or green) after also passing through a control volume (white or gold). Approaches to landing pad, gold-from-white; landings, cyan-from-gold; approaches to dark spot, green-from-white. Number of trajectories per condition: 44–1,288 (control), 228–1,815 (odour). Experiments were performed at two concentrations: 15 ml min⁻¹ (left) and 60 ml min⁻¹ (right). Letters above data indicate statistically significant groups (two-tailed Mann–Whitney *U*-test at *P* < 0.05 with eight-way Bonferroni corrections). In all panels, shading indicates the bootstrapped 95% confidence interval around the mean.

¹California Institute of Technology, Pasadena, CA, USA. ²Present address: University of Nevada, Reno, NV, USA. *e-mail: flyman@caltech.edu

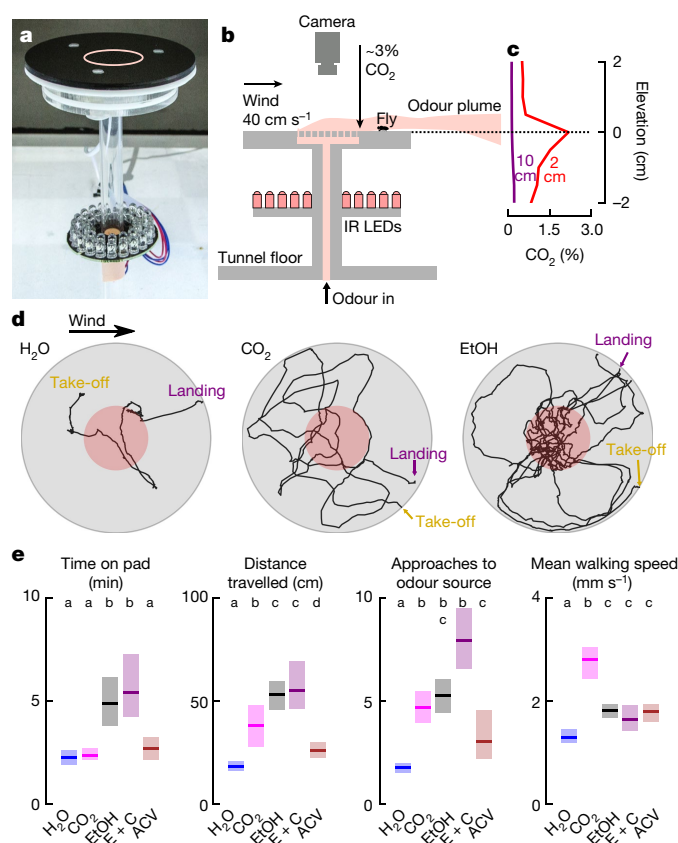


Fig. 2 | Walking *Drosophila* are attracted to CO₂. **a**, Photograph of landing platform. **b**, Cross-sectional diagram of the landing platform. **c**, CO₂ concentration for two altitude transects, 2-cm and 10-cm downwind from the platform, at a 60 ml min⁻¹ flow rate added to 100 ml min⁻¹ of clean air (Extended Data Fig. 2a, b). **d**, Stereotypical trajectories. **e**, Four descriptive statistics that summarize the behaviour of flies in response to different odours. Flow rate was 60 ml min⁻¹ for each odour added to 100 ml min⁻¹ of clean air. ACV, apple cider vinegar; E + C, 60 ml min⁻¹ clean air bubbled through ethanol with 15 ml min⁻¹ of CO₂ added. See Extended Data Fig. 2c for additional flow rates. Shading indicates the bootstrapped 95% confidence interval around the median. Number of trajectories = 125–193 per odour. Approaches to odour = number of times trajectories entered the red region shown in **d**. Letters above data indicate statistically significant groups (two-tailed Mann–Whitney *U*-test at *P* < 0.05 with five-way Bonferroni corrections).

the flight trajectories of flies in a wind tunnel that contained a landing platform, which was programmed to periodically release plumes of CO₂ or ethanol (Fig. 1a, b). Both odours elicited approaches, landings and explorations of a conspicuous visual feature (Fig. 1c, d), which is consistent with previous experiments with flies and mosquitoes^{15,16}. Flies were more likely to approach the platform or dark spot in the presence of ethanol compared to CO₂, but were equally likely to land in response to either odour (Fig. 1e).

To quantify the behaviour of flies after they land, we designed a platform that is suitable for automated tracking (Fig. 2a, b). At a flow rate of 60 ml min⁻¹ CO₂, the CO₂ concentration near the surface of the platform was approximately 3% (Fig. 2b, c). After landing near a source of CO₂, ethanol or apple cider vinegar, flies exhibited a local search behaviour that was similar to so-called ‘dances’¹⁷ (Fig. 2d, e, Extended Data Fig. 2a–c). Flies spent twice the amount of time exploring platforms that emitted ethanol compared to CO₂ or vinegar. Flies approached a source that emitted both ethanol and CO₂ more frequently than they approached vinegar, or either odour alone. Vinegar elicited smaller local searches and slightly fewer approaches compared to CO₂, consistent with the hypothesis that vinegar might indicate a less favourable, late-stage ferment. Flies spent significantly less time standing still on

the platform in the presence of CO₂ compared to any other odour, with a mean walking speed > 2 mm s⁻¹ (Fig. 2e).

One previous study showed that *Drosophila* flies are attracted to CO₂ while flying on a tether¹⁸. Our results confirm this observation in freely flying flies; however, we also found that flies remain attracted to CO₂ after they land, which contradicts previous studies^{7–10,12}. One potential explanation is that flies in constrained walking chambers might behave differently to those that arrived on our open wind tunnel platform after tracking the odour plume and landing. To test this hypothesis, we built an enclosed arena in which flies were unable to fly (Fig. 3a, Extended Data Fig. 3) and presented them with pulses of 5% CO₂. Groups of 10 starved flies presented with CO₂ after acclimating to the arena for 10 min exhibited aversion (Fig. 3b), as previously reported. However, if allowed to acclimate in the chamber for two hours, the flies exhibited attraction to CO₂ (Fig. 3c).

To study the response of these flies in more detail, we recorded the behaviour of flies for 20 h, while providing 10-min presentations of CO₂ from alternating sides of the arena every 40 min (Fig. 3d, Supplementary Videos 1, 2). To control for humidity, we continuously pumped 20 ml min⁻¹ of H₂O-saturated air through the odour ports on both sides of the chamber. The flies exhibited a clear circadian rhythm within the chamber, as indicated by their mean walking speed. At times of peak activity—near dusk and dawn—flies showed a strong initial attraction to CO₂, which decayed stereotypically during the 10-min presentation. At times of low activity—at mid-day and during the night—flies exhibited a mild aversion to CO₂. Starving flies for 24 h before the experiment changed their activity profile, resulting in a slightly elevated attraction during the night. Ethanol, by contrast, elicited sustained attraction regardless of baseline activity (Fig. 3d, Supplementary Video 3).

To probe this relationship between activity and CO₂ attraction, we increased the temperature and elevated the wind speed—manipulations that are known to elevate and depress¹⁹ activity, respectively (Fig. 3e). When we increased the bulk-flow rate to 100 ml min⁻¹, flies exhibited a peak walking speed of about 1.5 mm s⁻¹ at dusk—nearly half the speed we measured at a flow rate of 20 ml min⁻¹. Instead of showing attraction, these flies exhibited aversion to 5% CO₂, although they were still attracted to ethanol (Fig. 3e). This result helps to explain why previous studies that used higher flows (100–1,000 ml min⁻¹) to present CO₂ observed aversion⁸. To further explore the effect of wind, we clipped the arista of the flies, which destroys their primary means of detecting airflow but does not interfere with the detection of odours²⁰. The flies without arista exhibited the same walking speed and attraction to CO₂ at the high flow rate as was exhibited by normal flies at the low flow rate. Warming flies with intact arista to 32 °C also increased their baseline activity and recovered their attraction to CO₂ at the higher flow rate. Pooling data across all our experimental conditions, we found that flies were attracted to CO₂ when they had a baseline walking speed that was above about 2.4 mm s⁻¹ (Fig. 3f). This value is similar to the walking speed that we observed in our wind tunnel assay, which was higher for CO₂ than the other odours. To confirm that activity-dependent attraction to CO₂ is not a function of social interactions, we tested 29 single flies, which behaved similarly to the cohorts of 10 (Extended Data Fig. 4a). We also tested three concentrations of CO₂ (1.7%, 5% and 15%) and found that the 5% concentration elicited the strongest response, consistent with our wind tunnel experiments (Extended Data Fig. 4b–f, Supplementary Information).

Although the responses of flies to ethanol and CO₂ were similar at stimulus onset, attraction to ethanol was more sustained. The time course of behaviour was notably similar in the walking arena and wind tunnel (Extended Data Fig. 2d–g), which suggests that the behavioural dynamics of olfactory attraction are robust to the stimulus environment and may represent an adaptation for using information that broad (CO₂) and more specific (ethanol) odorants provide.

Previous research shows that CO₂ aversion is mediated by Gr63a and Gr21a receptors^{7,9,21}; high concentrations of CO₂ are also detected by an acid-sensitive ionotropic receptor, IR64a¹⁰. In our assay, mutant

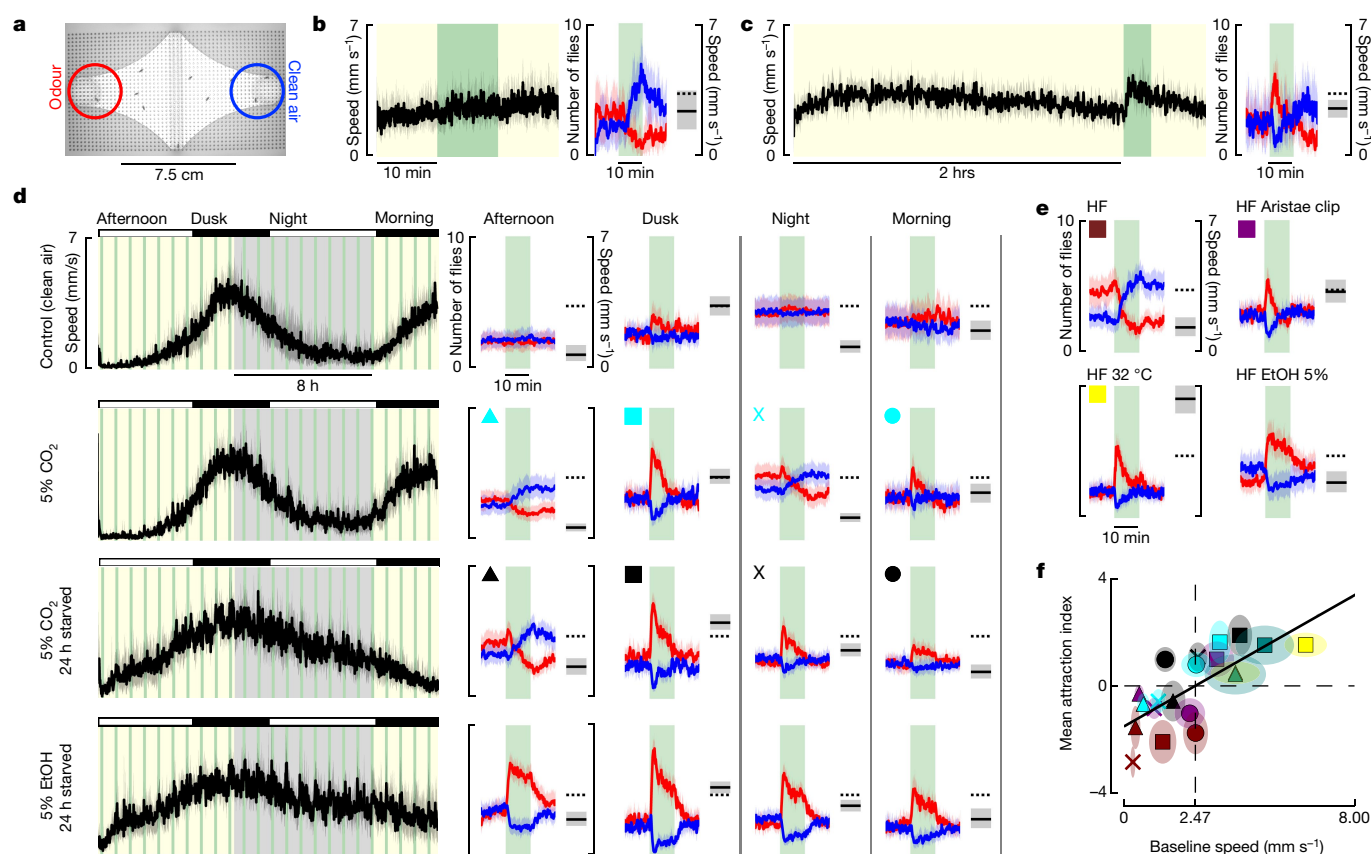


Fig. 3 | Attraction to CO₂, but not ethanol, depends on activity. **a**, Image of walking arena, with regions of interest near clean air (blue) and odour (red). **b**, Left, mean speed of 10 starved flies. Green, time at which 1 ml min⁻¹ of odour was added to 20 ml min⁻¹ bulk flow (from alternating sides). Right, number of flies in regions of interest near the CO₂ (red) and clean air (blue). Black bar and shading shows the mean speeds of flies 5 min before odour presentation, which is a proxy for activity level. $n = 8$ cohorts. **c**, Same as **b**, with 2-h acclimatization period. $n = 10$ cohorts. **d**, Left, mean speed of flies in a 20-h experiment. Shading indicates the entrained day (yellow) and night (grey) cycle. Right, data plotted as in **b**, for four time frames. For the control, we added clean air to the flow. Flies are significantly less attracted to this mechanical stimulus than they are to the olfactory ones. Dashed line, speed at dusk during clean air control (top row), also plotted in **b**, **c**, **e**. **e**, Manipulating the activity of flies changes their attraction to CO₂. Data are shown for experiments similar to those in **d** (dusk) but under 100 ml min⁻¹ bulk-flow conditions. In

these experiments, 5 ml min⁻¹ odour was added (same concentration as in **d**). Experiments were performed with intact flies (maroon), flies with aristae surgically removed (purple), and intact flies at 32° under a heat lamp (yellow). We also tested intact flies using an ethanol stimulus. HF, high flow. **f**, Summary of CO₂ responses presented in **d** and **e**, showing the relationship between activity and CO₂ attraction. Colour and shape encodes experiment and time of day (as shown in **d**, **e**). Green data are from experiments at 20 ml min⁻¹ bulk flow and 32°C. The mean attraction index represents the mean number of flies in regions of interest near CO₂ during stimulus, minus the number of flies in regions of interest 5 min before stimulus. Baseline speed, mean speed of all flies 5 min before CO₂ stimulus. Shading indicates the 95% confidence interval around the mean. All experimental combinations were performed with $n = 6$ cohorts of 10 flies each, and 24–48 trials per condition. Additional statistical analyses are provided in the Supplementary Information.

flies that lack the IR64a receptor showed no significant change in their behaviour compared to wild type (Fig. 4a, b, d). Consistent with previous work, mutants that lack the Gr63a receptor exhibited no aversion to CO₂; however, they were still attracted to CO₂ when active. Mutant flies that are homozygous for both Gr63a and IR64a behaved similarly to the Gr63a mutants. It is noteworthy that the characteristic decaying time course of attraction was unaffected in Gr63a mutants, even though these flies showed no aversion. Thus, the decay in attraction to CO₂ is not caused by an increase in aversion over time.

Given that CO₂ attraction is not mediated by Gr63a, Gr21a or IR64a, we wanted to confirm that the attraction is indeed a chemosensory response. To determine whether CO₂ attraction is mediated by either an olfactory or ionotropic receptor, we tested a mutant that lacks the olfactory and ionotropic co-receptors (Orco, IR25a and IR8a) as well as Gr63a (Fig. 4c). These near-anosmic mutants exhibited no detectable behavioural response to CO₂. Flies in which we surgically removed the third antennal segment also showed no response to CO₂, despite normal levels of activity. Together with our arista ablations (Fig. 3e), these experiments show that CO₂ attraction is mediated by receptors on the third antennal segment. To further confirm this, we tested each

co-receptor mutant individually and found that mutants that lack IR25a did not exhibit wild-type CO₂ attraction, whereas Orco and IR8a mutants did (Fig. 4c). Mutant flies that lack Orco, IR8a and Gr63a also exhibit wild-type attraction to CO₂, confirming that the only required co-receptor is IR25a. IR25a has previously been implicated in a wide range of behaviours, including temperature^{22,23} and humidity²³ sensation. We measured the temperature in our arena near the CO₂ port, and found no change in temperature as a result of the stimulus (Extended Data Fig. 5). To eliminate the possibility of a humidity artefact, we tested an IR40a mutant, which still exhibited attraction to CO₂ (Fig. 4c). In summary, our experiments show that CO₂ attraction is mediated by a separate chemosensory pathway from that which governs aversion, and that CO₂ attraction requires the IR25a co-receptor (Fig. 4d). IR25a is the most highly conserved olfactory receptor among insects^{24,25}. It is possible that other insect species that lack Gr63a²⁶ but that still respond to CO₂ use the same IR25a-dependent pathway. Unfortunately, the GAL4 driver for the IR25a promoter is expressed only in about half of the endogenous IR25a-expressing neurons²⁷, which makes imaging experiments that aim to identify which glomerulus is involved difficult at this time.

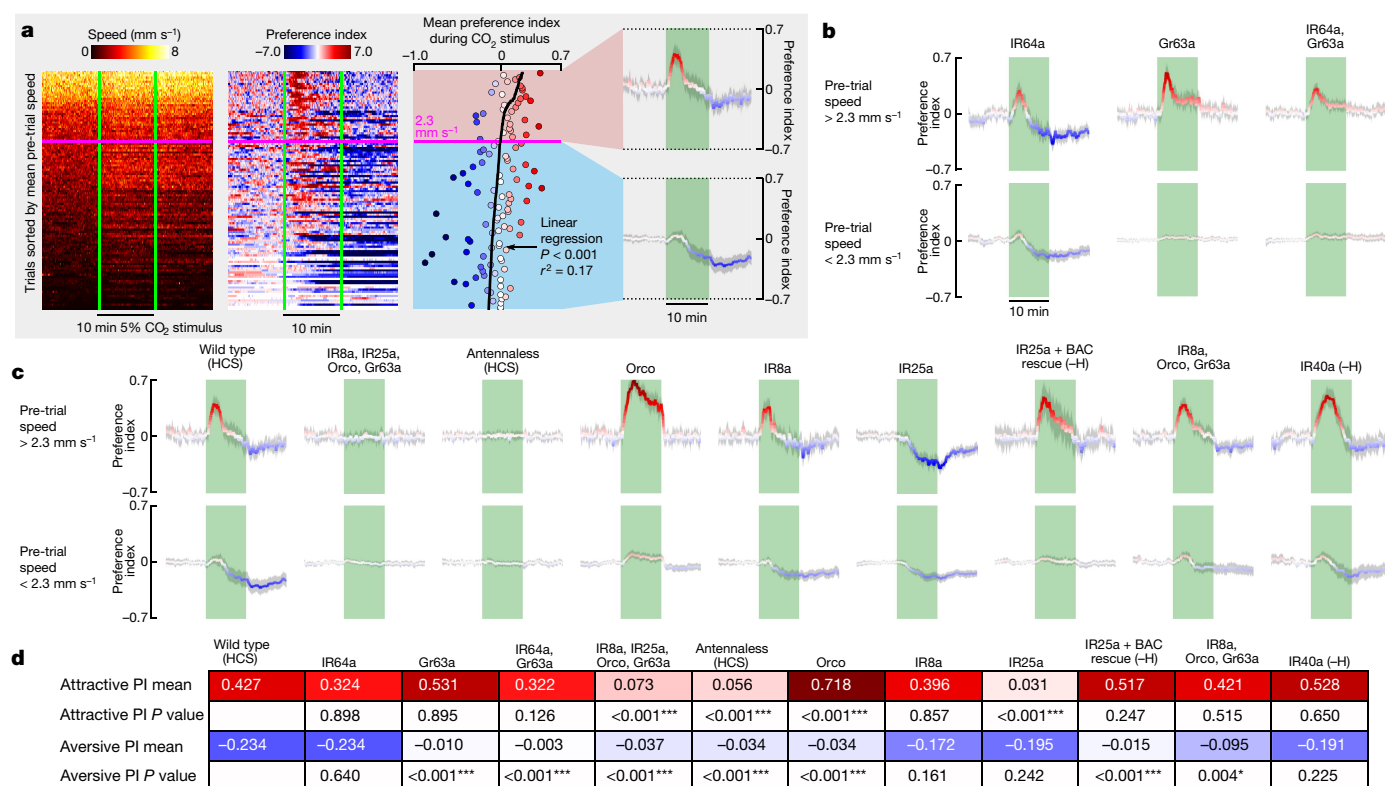


Fig. 4 | Attraction and aversion to CO₂ are mediated by separate chemosensory pathways. **a**, Data from 10 cohorts of flies with a 5% CO₂ stimulus sorted by the mean speed (S) during a reference period 5 min before stimulus presentation (S_{Ref}). To achieve a range of baseline activities, 4 cohorts were starved for 24 h, 3 cohorts were starved for 3 h and 3 cohorts were starved for 3 h and heated to 32°C, $n = 112$ trials. The preference index (PI) was calculated in two steps: (1) $PI_0 = (n_{\text{odour}} - n_{\text{control}})/n_{\text{total}}$ and (2) $PI = PI_0 - PI_0|_{\text{Ref}}$, in which n = number of flies, and $n_{\text{total}} = 10$ flies in total per cohort. We determined the linear regression for the mean preference index during the stimulus with respect to S_{Ref} , and used the intercept to cluster the data into high- and low-activity groups. For these groups, we calculated the mean preference index over time. **b**, **c**, Data plotted as in last panel of **a**, for different manipulations and mutants, using the intercept of 2.3 mm s⁻¹ found in **a** to cluster the data. All flies were presented with randomly

interleaved stimuli of 0% or 5% (5% responses are shown here, see Extended Data Fig. 6 for 0% responses). $n = 16$ –110 trials per condition. Shading indicates the bootstrapped 95% confidence interval around the mean for **a**–**c**. HCS, Heisenberg Canton-S stock. **d**, Summary of statistics for each mutant. Top row shows the mean largest preference index for the active group during the stimulus. Second row shows the mean smallest preference index for the inactive group during the stimulus. Third and fourth rows show the P values for a two-tailed Kolmogorov–Smirnov test between the mutant and wild type. Bonferroni-corrected statistically significant differences are indicated with asterisks. *** $P < 0.005$; * $P < 0.05$. BAC, bacterial artificial chromosome. For mutants followed by ‘-H’, we omitted the data collected at 32°C, because our analysis found these flies did not respond to CO₂ despite responding under more-natural 24-h-starved conditions (see Extended Data Fig. 7).

Our finding that active flies are attracted to CO₂ makes ethological sense, given that CO₂ is generated by yeast—the preferred food of these flies. We considered why it might be that *Drosophila* avoid CO₂ when in a low-activity state. Flies do not exhibit this state-dependent reaction to ethanol and vinegar (Extended Data Fig. 8); perhaps the aversion to CO₂ at low activity is an adaptation that minimizes encounters with parasites that seek CO₂. Alternatively, the behaviour may help flies to avoid respiratory acidosis when near high concentrations of CO₂ within the environment¹⁴ (Extended Data Fig. 9). Previous studies have suggested that CO₂ serves as an aversive pheromone by which stressed flies signal others to flee a local environment⁷. However, an alternative explanation is that agitated flies release CO₂ not as a social signal but simply because it is present in their tracheal system owing to their process of discontinuous respiration^{28,29} (Extended Data Fig. 10). Further work on this state-dependent reaction to CO₂ will require experiments that carefully consider the natural ethology of the flies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0732-8>.

Received: 22 December 2017; Accepted: 11 October 2018;

Published online 21 November 2018.

- Guerenstein, P. G. & Hildebrand, J. G. Roles and effects of environmental carbon dioxide in insect life. *Annu. Rev. Entomol.* **53**, 161–178 (2008).
- Dekker, T. & Cardé, R. T. Moment-to-moment flight manoeuvres of the female yellow fever mosquito (*Aedes aegypti* L.) in response to plumes of carbon dioxide and human skin odour. *J. Exp. Biol.* **214**, 3480–3494 (2011).
- Thom, C., Guerenstein, P. G., Mechaber, W. L. & Hildebrand, J. G. Floral CO₂ reveals flower profitability to moths. *J. Chem. Ecol.* **30**, 1285–1288 (2004).
- Buehlmann, C., Hansson, B. S. & Knaden, M. Path integration controls nest-plume following in desert ants. *Curr. Biol.* **22**, 645–649 (2012).
- Stange, G. Carbon dioxide is a close-range oviposition attractant in the Queensland fruit fly *Bactrocera tryoni*. *Naturwissenschaften* **86**, 190–192 (1999).
- Klocke, D., Schmitz, A. & Schmitz, H. *Native Flies Attracted to Bushfires* (Department of Environment and Conservation, The Government of Western Australia, 2009).
- Suh, G. S. B. et al. A single population of olfactory sensory neurons mediates an innate avoidance behaviour in *Drosophila*. *Nature* **431**, 854–859 (2004).
- Faucher, C., Forstreuter, M., Hilker, M. & de Bruyne, M. Behavioral responses of *Drosophila* to biogenic levels of carbon dioxide depend on life-stage, sex and olfactory context. *J. Exp. Biol.* **209**, 2739–2748 (2006).
- Jones, W. D., Cayirlioglu, P., Kadow, I. G. & Vosshall, L. B. Two chemosensory receptors together mediate carbon dioxide detection in *Drosophila*. *Nature* **445**, 86–90 (2007).
- Ai, M. et al. Acid sensing by the *Drosophila* olfactory system. *Nature* **468**, 691–695 (2010).
- Faucher, C. P., Hilker, M. & de Bruyne, M. Interactions of carbon dioxide and food odours in *Drosophila*: olfactory hedonics and sensory neuron properties. *PLoS ONE* **8**, e56361 (2013).
- Lin, H., Chu, L., Fu, T., Dickson, B. J. & Chiang, A. Parallel neural pathways mediate CO₂ avoidance responses in *Drosophila*. *Science* **340**, 1338–1341 (2013).

13. Sorey, M. L. et al. *Invisible CO₂ Gas Killing Trees at Mammoth Mountain, California* US Geological Survey Fact Sheet 172-96 (USGS, 2000).
14. Hubbard, H. G. Insect life in the hot springs of Yellowstone National Park. *Can. Entomol.* **23**, 226–235 (1891).
15. van Breugel, F., Riffell, J., Fairhall, A. & Dickinson, M. H. Mosquitoes use vision to associate odor plumes with thermal targets. *Curr. Biol.* **25**, 2123–2129 (2015).
16. van Breugel, F. & Dickinson, M. H. Plume-tracking behavior of flying *Drosophila* emerges from a set of distinct sensory-motor reflexes. *Curr. Biol.* **24**, 274–286 (2014).
17. Kim, I. S. & Dickinson, M. H. Idiothetic path integration in the fruit fly *Drosophila melanogaster*. *Curr. Biol.* **27**, 2227–2238 (2017).
18. Wasserman, S., Salomon, A. & Frye, M. A. *Drosophila* tracks carbon dioxide in flight. *Curr. Biol.* **23**, 301–306 (2013).
19. Yorozu, S. et al. Distinct sensory representations of wind and near-field sound in the *Drosophila* brain. *Nature* **458**, 201–205 (2009).
20. Gaudry, Q., Nagel, K. I. & Wilson, R. I. Smelling on the fly: sensory cues and strategies for olfactory navigation in *Drosophila*. *Curr. Opin. Neurobiol.* **22**, 216–222 (2012).
21. Kwon, J. Y., Dahanukar, A., Weiss, L. A. & Carlson, J. R. The molecular basis of CO₂ reception in *Drosophila*. *Proc. Natl Acad. Sci. USA* **104**, 3574–3578 (2007).
22. Ni, L. et al. The ionotropic receptors IR21a and IR25a mediate cool sensing in *Drosophila*. *eLife* **5**, e13254 (2016).
23. Enjin, A. et al. Humidity sensing in *Drosophila*. *Curr. Biol.* **26**, 1352–1358 (2016).
24. Silbering, A. F. et al. Complementary function and integrated wiring of the evolutionarily distinct *Drosophila* olfactory subsystems. *J. Neurosci.* **31**, 13357–13375 (2011).
25. Croset, V. et al. Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. *PLoS Genet.* **6**, e1001064 (2010).
26. Robertson, H. M. & Kent, L. B. Evolution of the gene lineage encoding the carbon dioxide receptor in insects. *J. Insect Sci.* **9**, 19 (2009).
27. Abuin, L. et al. Functional architecture of olfactory ionotropic glutamate receptors. *Neuron* **69**, 44–60 (2011).
28. Lighton, J. R. B. Discontinuous gas exchange in insects. *Annu. Rev. Entomol.* **41**, 309–324 (1996).
29. Hetz, S. K. & Bradley, T. J. Insects breathe discontinuously to avoid oxygen toxicity. *Nature* **433**, 516–519 (2005).

Acknowledgements We thank A. Straw for the 3D tracking software. Several colleagues provided mutants: R. Benton (quadruple mutant), R. Stanewsky (IR25a and rescue); G. Suh (IR8a); and M. Gallio and M. Stensmyr (IR40a). R. Benton, E. Hong and J. Riffell contributed helpful comments. This work was funded by grants from NIH (NIH1R01DC013693-01, U01NS090514) and the Simons Foundation.

Reviewer information *Nature* thanks S. Combes, M. Frye, L. Vosshall and R. Wilson for their contribution to the peer review of this work.

Author contributions F.v.B. and M.H.D. conceived the experiments. A.H. made genetic recombinants. F.v.B. and A.H. performed experiments. F.v.B. analysed data. F.v.B. and M.H.D. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0732-8>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0732-8>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.H.D.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Statistics and reproducibility. Here we provide the exact number of trials, trajectories, individuals and cohorts for each experiment. For our wind tunnel experiments, each trajectory was treated as an independent sample because it is impossible to keep track of the identity of individual flies in these experiments. In the walking assays, each trial was considered independent, as the inter-trial variability within a cohort of flies over the course of the 20-h experiments was similar to the inter-cohort variability. This is in part due to the changes in activity over the course of the experiments. In all of our figures, we show the trial by trial variance with shaded 95% confidence intervals around the mean or median. These confidence intervals were determined by 1,000 iterations of bootstrapped sampling with replacement. In each experiment, we attempted to collect the largest sample sizes we could, given the time constraints required for behavioural data in which an experiment with one cohort lasts for 24 h. In situations in which we were comparing behaviour under different conditions, we attempted to randomize the temporal sequence with which we collected data to minimize any artefacts due to long-term influences such as season changes in humidity, temperature and so on. We did not use blinding in our data collection design. For experiments described in Figs. 1, 4, and the associated Extended Data Figs., control and test experiments were interleaved with each other.

Additional statistics for Fig. 3. To statistically compare the attraction of flies to CO₂ under the different conditions presented in Fig. 3, we used resampling (Fisher's exact test) to test the significance of the difference in the preference indices exhibited by flies in key experiments. The preference index represents the strength of the flies' attraction to the odour (for example, CO₂), relative to the clean air control. The raw preference index, PI_0 , was first calculated for each point in time: $PI_0(t) = (n_{\text{odour}}(t) - n_{\text{control}}(t))/n_{\text{total}}$, in which n_{odour} and n_{control} are the number of flies within the circular regions of interest around the odour and control ports, respectively, and n_{total} is the total number of flies. To remove baseline biases, we then subtracted the mean preference index for the 5-min period before the odour stimulus to yield the relative preference index, PI :

$$PI(t) = PI_0(t) - \overline{PI_0(t)}_{t=-5:0}$$

To make statistical comparisons, we then calculated the average preference index for the first half of the odour presentation period (that is, the first 5 min). We chose this range because it captures the majority of CO₂ attraction, and thus focuses the statistical test on the most relevant time period.

$$\overline{PI_{t=0:5}} = \overline{PI(t)}_{t=0:5}$$

These calculations provide a single preference index value for each trial of each cohort. For our resampling algorithm, we used 1,000 iterations to determine the P value, and repeated this calculation 1,000 times to calculate a 95% confidence interval around these P values. The confidence intervals are shown for key comparisons as follows. Figure 3b compared to Fig. 3c: $0.0249 < P \text{ value} < 0.0276$. Figure 3d (top row) dusk compared to Fig. 3d (top row) afternoon: $P \text{ value} = 0.002$. Figure 3d (top row) dusk compared to Fig. 3d (second row) dusk: $0 < P \text{ value} < 0.001$.

To compare the response of flies to CO₂ and ethanol, we used the full 10-min odour-presentation time frame because the differences in behaviour primarily appear in the second half of the odour presentation. Figure 3d (third row) dusk compared to Fig. 3d (fourth row) dusk: $0 < P \text{ value} < 0.001$ (24-h-starved flies). Extended Data Fig. 2e compared to Extended Data Fig. 2g (red traces): $0 < P \text{ value} < 0.001$ (12-h-starved flies).

To eliminate the possibility of pseudo-replication, we repeated our statistics after calculating the average $PI(t)$ for each cohort before calculating $\overline{PI_{t=0:5}}$. Thus, for the following statistics, the input to our resampling test was a single preference index value for each cohort of flies. This is a very conservative measure, because there is similar intra-cohort variability compared to inter-cohort variability, in part owing to changes in the flies' circadian activity. Figure 3b compared to Fig. 3c: $0.0249 < P \text{ value} < 0.0276$ (these experiments were 1 trial per cohort). Figure 3d (top row) dusk compared to Fig. 3d (top row) afternoon: $0.0124 < P \text{ value} < 0.0141$. Figure 3d (top row) dusk compared to Fig. 3d (second row) dusk: $0.0129 < P \text{ value} < 0.0149$. Figure 3d (third row) dusk compared to Fig. 3d (fourth row) dusk: $0.0100 < P \text{ value} < 0.0120$ (24-h-starved flies). Extended Data Fig. 2e compared to Extended Data Fig. 2g (red traces): $0.0035 < P \text{ value} < 0.0045$ (12-h-starved flies).

This definition of preference index was also used for the data presented in Fig. 4. **Flies.** Wild-type flies were descendants of a Heisenberg Canton-S stock (HCS). For the arista-clipped and antennaless flies, we cold-anaesthetized flies and carefully removed the arista or third antennal segment with sharpened forceps.

Each mutant used in our study is described in detail below. All experiments were done with mutants in which balancers and markers had been crossed out.

Gr63a, IR64a: +/+; Gr63a^{-/-}IR64a^{-/-} double mutant; this line was generated using recombination by crossing $TI\{w^{+m*} = TI\}Gr63a^1$ (Bloomington 9941) to $Mi\{ET1\}Ir64a^{MB05283}$ (Bloomington 24610). The double mutants were verified using PCR. IR8a, IR25a, Orco, Gr63a (near-anosmic): $IR8a^{-/-}IR25a^{-/-}Orco^{-/-}Gr63a^{-/-}$ quadruple mutant; this was a gift from R. Benton and A. Silberling³⁰. IR8a, Orco, Gr63a: $IR8a^{-/-};+;Orco^{-/-}Gr63a^{-/-}$ triple mutant; this line was generated by crossing IR8a;IR25a;Orco,Gr63a to wild-type HCS. Orco: +/+;Orco^{-/-}; this line was created by backcrossing an Orco² (Bloomington 23130) line to the wild-type HCS for five generations, and verified through PCR. IR64a: +/+; IR64a^{-/-}; this line was created by backcrossing the $Mi\{ET1\}Ir64a^{MB05283}$ (Bloomington 24610) line to the wild-type HCS for seven generations, and verified through PCR. IR8a: $IR8a^{-/-};+;+$; this mutant was a gift from G. Suh^{27,31}. IR25a: +/+;IR25a^{-/-};+; we used two variants of this mutant ((1) and (2)), along with the bacterial artificial chromosome rescue, all of which were gifts from R. Stanewsky. Figure 4 uses the (2) variant. Gr63a: +/+;Gr63a^{-/-}; this mutant is Bloomington 9941. IR40a: +/+;IR40a^{-/-};+; this mutant was a gift from M. Stensmyr and M. Galio²³.

All of the flies were raised on a 16:8 light:dark light cycle at 25°C in standard 300-ml bottles on fly food consisting of: water (17.8 l), agar (136 g), cornmeal (1,335.4 g), yeast (540 g), sucrose (320 g), molasses (1.64 l), CaCl₂ (12.5 g), sodium tartrate (150 g), tegosept (18.45 g), 95% ethanol (153.3 ml) and propionic acid (91.5 ml). For all of our experiments, we used 2- to 3-day-old female flies. To sort and starve flies, they were briefly anaesthetized on a cold plate, and placed in a test-tube with a wet Kimwipe.

Fermentation and trap assays. We prepared the wort from 130 ml of apple juice (Treetop brand) and 20 g of cane sugar, warmed to 35°C. Next, we added 130 mg of Cellar Science EC-1118 wine yeast, which produces a neutral flavour and aroma. The fermentation was carried out at room temperature (23°C), under an airtight. All glassware was first sanitized with StarSan. We measured the specific gravity daily with a standard hydrometer, and calculated the alcohol content according to the following equation³²,

$$ABV = \left(76.08 \times \frac{OG - FG}{1.775 - OG} \right) \times \left(\frac{FG}{0.794} \right) \%$$

in which ABV is alcohol by volume, OG is the starting specific gravity and FG is the final specific gravity. After 14 days, the fermentation had finished and the yeast flocculated. At this point, we sealed the containers and stored them in the fridge for 6–14 days while waiting for the next active batch of ferments to reach the desired age.

For the trap assays we let fermentations run for 2, 7, or 12 days. One day before these ferments were ready, we pulled a flocculated ferment from the fridge, and wet-starved groups of flies (50–150 flies each). The following day we ran three trap assay trials. For each trial we poured the active ferment into one jar, and the flocculated ferment into another jar, and inserted the traps into the jars. The two traps were placed side-by-side in our wind tunnel (~6 cm apart), and a group of flies was released. Two hours later we removed the traps, CO₂-anaesthetized the flies, and counted the number of individuals in each trap. A preference index was calculated as: $(n_a - n_t)/(n_a + n_t)$, in which n_a is the number of flies in the active ferment, and n_t is the number of flies in the flocculated ferment. For each condition we used four separate ferments, each used for three separate trials, for a total of 12 trials per condition.

CO₂ measurements of fly bottles. We first modified 500-ml Nalgene bottles by drilling two holes and fitting them with Luer Lock valves (with lock plugs attached). These Nalgene bottles are slightly larger than standard (300-ml) food bottles used by many *Drosophila* laboratories, and can be fitted with the same standard-sized cotton plugs. For each Nalgene bottle, we melted the food from 1 fly food bottle (50 ml) in the microwave, and poured it inside. Once cooled, we added a measured amount of baker's yeast, depending on the experiment, and fitted the bottle with a cotton plug and placed it in a 25°C incubator for 2 days. For experiments with flies, we added 10 females and 15 males to each bottle and allowed them to lay eggs in the bottles for two days. Fourteen days later (when the majority of the flies had eclosed, and were ~2 days old), we made our measurements.

To measure the CO₂ content, we first pressed the cotton plug into the bottle far enough to twist on the original Nalgene cap, sealing the contents of the bottle inside. Meanwhile, we prepared our CO₂ analyser—the LiCorr-6262—by running CO₂ free air through the system at 20 l min⁻¹.

We attached one of the Luer valves on the Nalgene to the input of the CO₂ analyser. Next, we quickly attached the CO₂-free air stream to the other Luer valve, slowly replacing the air inside the bottle with CO₂-free air. Before connecting the air stream, we started our data acquisition. Data were collected from the LiCorr-6262 using the analogue-to-digital converters on a Phidgets InterfaceKit, connected to an Ubuntu laptop running custom Python code for data acquisition.

Preliminary measurements showed that the CO₂ content of the bottles was beyond the dynamic range of the LiCorr-6262. To resolve this, we added a 500-

ml container filled with CO₂-free air as a buffer between the Nalgene bottle and the LiCorr. This buffer had the effect of spreading the CO₂ content over a longer time frame, reducing the concentration, which enabled us to accurately measure it. This approach, however, does not provide a direct measure of the CO₂ concentration. For this, we performed a calibration by filling the 500-ml Nalgene bottles with air of a known CO₂ concentration, and performing the experiment with these calibration bottles. After calibrating with three separate concentrations of 400, 2,000, and 10,000 p.p.m., we found a linear relationship between our measured peak CO₂ concentration, and the actual concentration of the bottles. Using this calibration curve, we were able to calculate the actual CO₂ concentration of the Nalgene bottles filled with fly food based on their measured peak CO₂ concentrations.

Free-flight wind tunnel assays. To record the free-flight behaviour of flying flies, we used the same wind tunnel and 3D tracking system described at length in previous papers^{16,33,34}. To observe the flies' behaviour in response to odours, we added an acrylic platform with two sites for odour release. Air flow was controlled using computer-controlled Alicat mass-flow controllers (0–200 ml min^{−1} range). For these and all other experiments, we used Teflon tubing. Cohorts of 12 female flies were starved for 6 h before starting the experiments at 17:00, 6 h before the flies' sunset. Starting at 20:00 (3 h before the flies' sunset), either CO₂ or ethanol was released from the landing platform for 30 min, followed by an hour of clean air. This stimulus pattern was repeated seven times.

Regions of interest. We chose regions of interest to quantify the behaviour of the trajectories shown in the heat maps of Fig. 1c, d. The boundaries of the regions for approaching the dark spot, approaching the platform, and landing on the platform were chosen based on the behaviour of the flies in the presence of the odours. The objective was to compare the behaviour with the different odours and controls, rather than determine absolute numbers. Thus, the exact size and position of the regions is not critical.

The white region of interest was chosen to be roughly in the region in which the odour plume passes, above and behind the dark spot. By comparing how many flies approach the pad or spot to how many flies pass through this white region, we control for the overall change in behaviour of the flies in the presence of the odour. For example, it is possible that the odour causes the flies to spend less time near the top of the tunnel, bringing them closer to the spot or platform—and thus more likely to approach these objects. By always selecting trajectories that passed through the same volume, we control for this overall change in behaviour.

Free-walking wind tunnel assays. The 3D tracking system used for the free-flight experiments did not have sufficient spatial and temporal resolution to accurately record the walking behaviour of flies once they had landed on the pad. To examine this behaviour more closely, we developed a 2D real-time tracking system designed for general-purpose applications. Our Python-based software and documentation is freely available on GitHub: http://florisvb.github.io/multi_tracker/. The software runs on Ubuntu, and is built on the ROS (Robot Operating System) framework, and takes advantage of open-source packages including OpenCV, scipy, numpy, pandas, h5py and pyQTgraph. A brief overview of the software flow is as follows: (1) image background subtraction; (2) thresholding and contour identification; (3) contours larger than a specified size are broken up into smaller contours (this corrects for cases when two flies come close to one another); (4) data association using a posteriori estimates from a Kalman filter estimator; (5) Kalman filtering of trajectories to (a) smooth position information, (b) estimate velocity and (c) calculate a posteriori estimates for the next data-association step; (6) trajectory data are recorded as an hdf5 file, and the changes from the background in the raw image are recorded as a ROS bag file; and (7) data can then be efficiently analysed using the pandas data structure, and trajectories can be viewed and corrected using a custom pyQTgraph GUI.

CO₂ plume measurements in the wind tunnel. We measured the CO₂ concentration downwind from the landing platform shown in Fig. 2a using a LiCorr-6262. To make accurate point measurements within the plume, we used a 15-cm-long tube with a 1-mm inner radius to minimize disturbances to the airflow. With a bulk air speed of 40 cm s^{−1}, the volume flow rate across the cross section of the tube was approximately 75 cm³ min^{−1} (ml min^{−1}). We used a mass-flow controller to regulate the suction being passed through the LiCorr-6262 to match this volume flow. After positioning the tube, we let the system equilibrate for several minutes before making a 2-min-long recording of the CO₂ concentration.

Because the LiCorr-6262 has a measurement limit of approximately 3,000 p.p.m. (0.3%), we made our measurements at low CO₂ flow rates (1–5 ml min^{−1}), and used a linear model to calculate the CO₂ concentration at larger flow rates (Extended Data Fig. 2a).

To further confirm our extrapolated measurements, we estimated the CO₂ concentration on the platform from first principles, as follows. First, we assume that all of the CO₂ that enters the wind tunnel is whisked away inside of the boundary layer (Extended Data Fig. 2b). The thickness of the boundary layer can therefore be used to estimate the average CO₂ concentration within that layer. The

thickness of the boundary layer can be approximated for laminar and turbulent flows as:

$$\delta_{\text{laminar}} = \frac{5x}{\sqrt{\text{Re}}} ; \delta_{\text{turbulent}} = \frac{0.37x}{\text{Re}^{1/5}}$$

in which δ is the thickness of the boundary layer, x is the distance downwind from the start of the platform and Re is the Reynolds number. With a characteristic length of 9 cm, a kinematic viscosity of $15 \times 10^{-6} \text{ m}^2 \text{ s}^{-1}$ for air at 20°C, and a free-stream velocity of 0.4 m s^{-1} , the Reynolds number is 2,400. For a value of $x = 6 \text{ cm}$, the boundary layers for laminar and turbulent flows are 6.1 mm and 4.6 mm, respectively. For simplicity, we will continue our calculations with a boundary layer of 5 mm.

The total volume flow rate over the platform can now be calculated as follows. The mean velocity in the boundary layer is 0.2 m s^{-1} (half the free-stream velocity), the CO₂ is released from a $3 \text{ cm} \times 3 \text{ cm}$ patch, and the boundary layer is 5-mm thick; thus, the total volume flow rate of clean air over the platform that is mixed with the introduced CO₂ is approximately $0.2 \times 0.03 \times 0.005 = 0.00003 \text{ m}^3 \text{ s}^{-1}$, or $1,800 \text{ ml min}^{-1}$. With 60 ml min^{-1} of CO₂ added, the concentration comes to 3.2%, which agrees relatively closely with our measurement model.

Walking assays. We designed custom walking arenas from sheets of laser-cut acrylic (Extended Data Fig. 3). Before experiments, the cut acrylic was washed with soap (Liquinox) and warm water, and wiped down with ethanol. Between each experiment, the floor and ceiling of the arenas were wiped down with ethanol. All walking experiments were done in darkness. Experiments for Fig. 3b, c were done during flies' peak activity (within 2 h of their subjective dusk).

Odour control in walking assays. While conducting experiments at low bulk-flow rates, we found that flies are very sensitive to minute changes in air flow, pressure and humidity. In an attempt to minimize the effect of these factors, we used three different stimulus architectures (Extended Data Fig. 3), all of which provided consistent results. The odours were controlled using a combination of computer controlled Alicat mass flow controllers and solenoid valves. Our ROS-based Python control software is available on GitHub at https://github.com/florisvb/multi_alicat_control. We used three different odour delivery architectures for our experiments as detailed below.

High flow. For our high flow (100 ml min^{-1} bulk-flow rate) experiments, we bubbled the 100 ml min^{-1} flow through MilliQ water, and added 5 ml min^{-1} clean dry air, CO₂ or clean dry air passed over a liquid ethanol reservoir, to the bulk flow. As a result of this architecture, during the odour presentation the flow rate of one side was slightly increased. However, experiments with clean dry air indicated that the flies did not respond to this change in flow rate. This arrangement was used for Fig. 4e.

Low flow, constant flow rate and humidity. At low flow rates (20 ml min^{-1} bulk-flow rate), the architecture used for the high-flow experiments did not work properly, as flies were attracted to the change in the overall flow rate. To overcome this, we re-designed the flow architecture. In this new system, we used additional mass-flow controllers that added 1 ml min^{-1} of clean dry air to the bulk-flow rate. During odour presentations, we used a solenoid to switch from 1 ml min^{-1} of clean dry air to CO₂, or clean dry air passed over liquid ethanol. This architecture ensured that the flow rate and the humidity on the two sides remained equal and constant. Control experiments in which we added clean dry air instead of CO₂ or ethanol confirmed that wild-type flies had minimal responses to the changes in flow. This arrangement was used for Fig. 3d.

Low flow, symmetric stimulus. The architecture used above (for 'Low flow, constant flow rate and humidity') elicited small responses in certain olfactory mutants. To achieve a complete null response in these flies, we re-designed the experimental architecture once more. In this third architecture, we removed the solenoids from the system because the flow transients they created appeared to be responsible for the responses of mutant flies. Instead, we connected two flow controllers to 20 ml min^{-1} bulk-flow lines. One of these flow controllers provided clean dry air, and the other CO₂. Both flow controllers were set to zero as a baseline. For each odour presentation, we added 3 ml min^{-1} of flow to both sides of the arena. One side received 3 ml min^{-1} of clean dry air, whereas the other received 2 ml min^{-1} of clean dry air and 1 ml min^{-1} of CO₂. In this arrangement, the flies experienced a change in the flow rate during odour presentations; however, the changes were symmetric. Furthermore, this arrangement made it possible to test different CO₂ concentrations ranging from 0% to ~15% on the same cohort of flies, providing continuous internal controls for our experiments. For these experiments, we reduced gain of the PID control settings on the mass flow controllers to provide smooth, slow change in flow rates. This is probably the cause of the slightly delayed behavioural responses that we observed. This arrangement was used for Figs. 3b–c, 4a–d, and Extended Data Figs. 4–8.

The qualitative—and even, to a large extent, quantitative—results across all three paradigms were consistent: at low activity the flies found CO₂ aversive, whereas at

high levels of activity the flies found CO₂ attractive. Finding the same results while working with three different olfactory-presentation architectures provides support for the robustness of our results.

Our experience with flies' sensitivity to changes in flow conditions—in particular at low bulk-flow rates—underscores how sensitive these flies are to odours and flow. Even our low flow rates of 20 ml min⁻¹ are reasonably high relative to the natural flow rates that a fly might experience on the surface or in the cracks of rotten fruit in the wild. The substantial changes in behaviour that we observed by reducing the flow rates to those better approximating field conditions highlights how important it is to consider the natural environments when studying sensory processing.

Temperature measurements. To eliminate any potential temperature-related confounds in our walking experiments, we measured the temperature in the arena near the odour ports using a thermistor rated to ± 0.1 °C (Omega brand model number 44031), connected to a Phidgets RTD sensor. Although we detected very small fluctuations in temperature throughout the day, we did not measure any changes in temperature that correlated with the presentation of our CO₂ stimulus (Extended Data Fig. 5).

Flies' fatal attraction to CO₂. During experiments with a 200 ml min⁻¹ CO₂ stimulus in the wind tunnel, some flies that approached the CO₂ were knocked out, as they would be on a typical CO₂ pad that is commonly used for sorting flies (Extended Data Fig. 9). While the average concentration of CO₂ just downwind from the odour stimulus would not have been lethal (10%, following the calculations associated with Extended Data Fig. 2a), the concentration right at the holes in the platform was 66% (200 ml min⁻¹ CO₂ added to 100 ml min⁻¹ of clean air).

CO₂ measurements of shaken insects. To measure the CO₂ produced by flies and mosquitoes when shaken in a vial, we placed 10–20 flies in a vial and pumped 100 ml min⁻¹ of CO₂-free air through the container. After 1 min, we forcefully tapped the vial against the table for 30 s, and measured the concentration of CO₂ in the air leaving the container using a LiCorr-6262. See Extended Data Fig. 10.

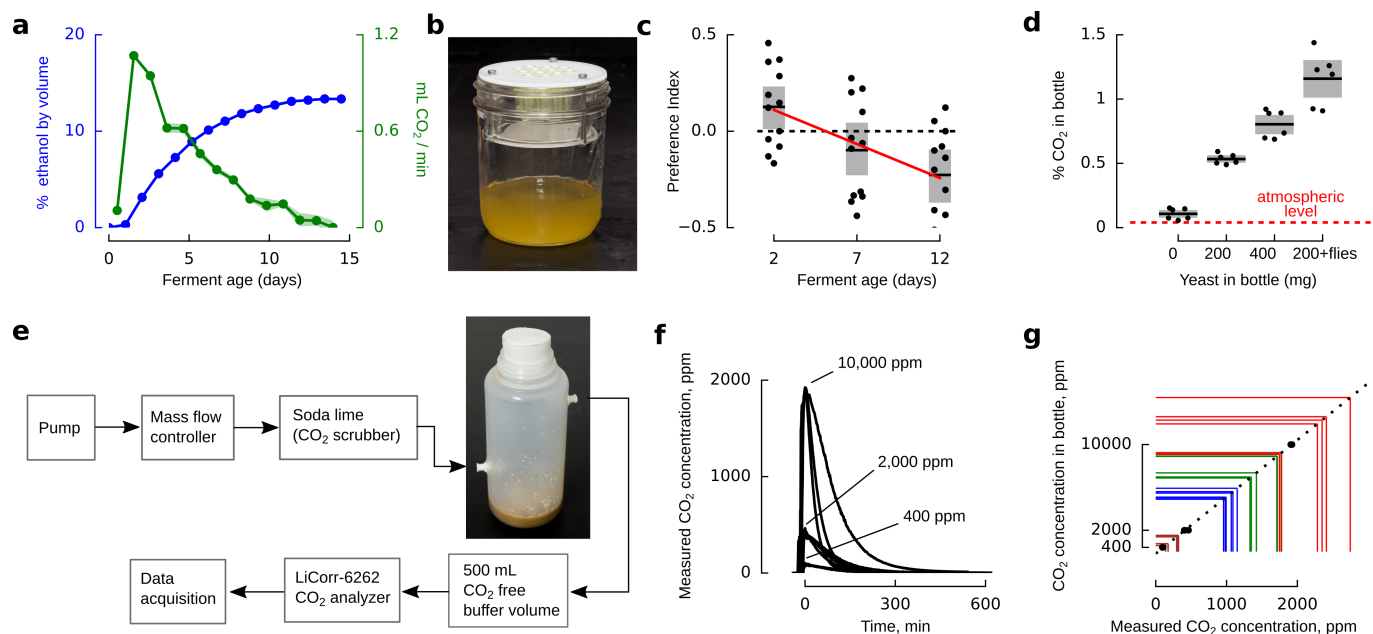
Code availability. Custom code is available online at https://github.com/florisvb/drosophila_co2_attraction.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

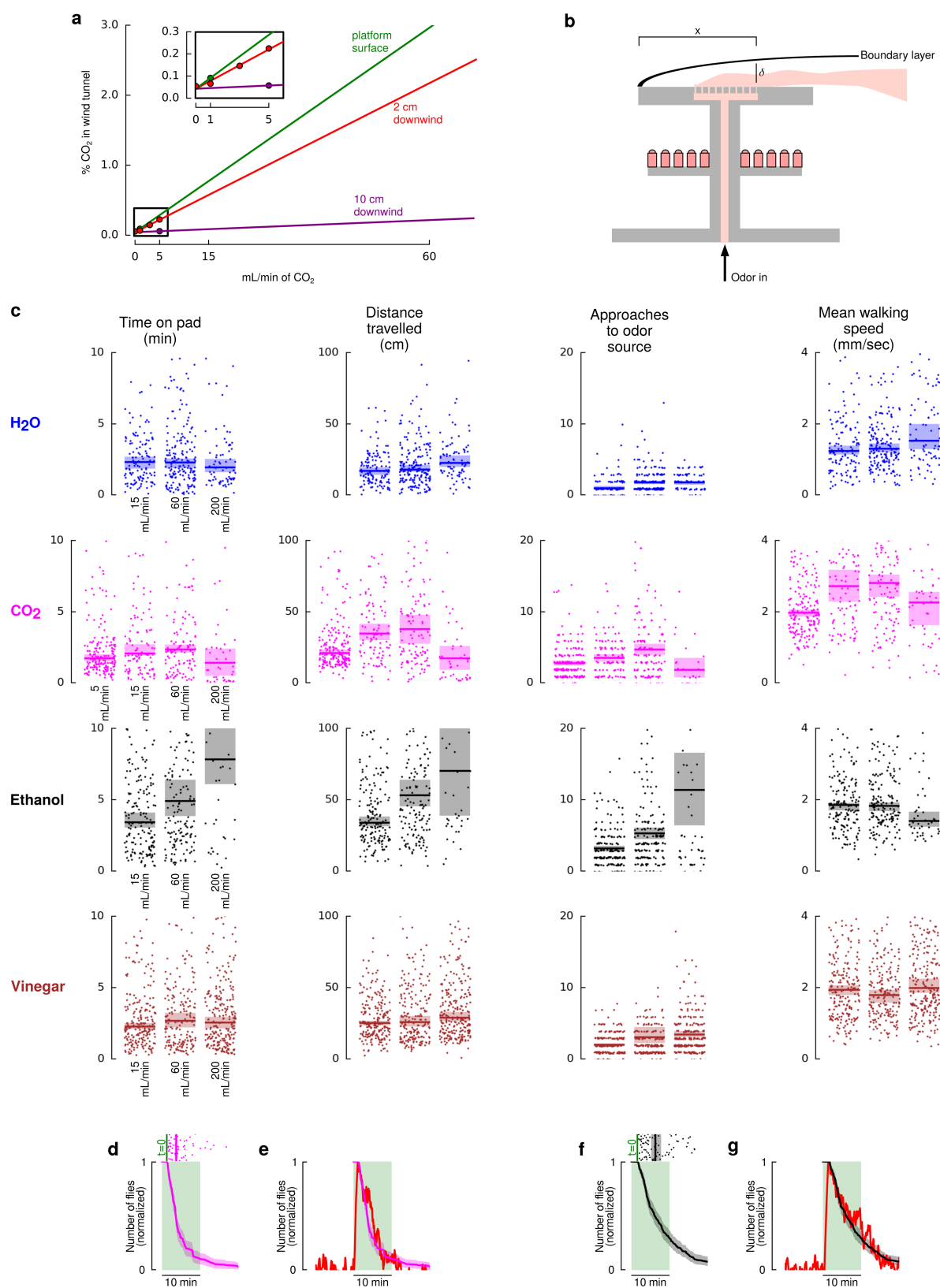
Processed data are available in a Dryad repository at <https://doi.org/10.5061/dryad.2s8422f>. Raw data are available from the corresponding author upon request.

30. Ramdya, P. et al. Mechanosensory interactions drive collective behaviour in *Drosophila*. *Nature* **519**, 233–236 (2015).
31. Ai, M. et al. Ionotropic glutamate receptors IR64a and IR8a form a functional odorant receptor complex in vivo in *Drosophila*. *J. Neurosci.* **33**, 10741–10749 (2013).
32. Hall, M. L. Brew by the numbers: add up what's in your beer. *Zymurgy* **1995-01-01**, 54–61 (1995).
33. Straw, A. D., Branson, K., Neumann, T. R. & Dickinson, M. H. Multi-camera real-time three-dimensional tracking of multiple flying animals. *J. R. Soc. Interface* **8**, 395–409 (2011).
34. Stowers, J. R. et al. Virtual reality for freely moving animals. *Nat. Methods* **14**, 995–1002 (2017).



Extended Data Fig. 1 | *Drosophila* prefer early fermentations, at peak CO₂ production. **a**, Alcohol by volume for apple juice and sugar fermented with champagne yeast over the course of two weeks, measured with a hydrometer. CO₂ production was calculated from the stoichiometry of fermentation (1 sugar molecule yields 2 ethanol and 2 CO₂ molecules), corresponding to the derivative of alcohol by volume. $n = 4$ independent ferments; the results were very consistent. **b**, Trap assay. **c**, Preference index exhibited by flies in three two-choice assays, using traps shown in **b**. Flies were presented with two traps: one was a completed 14-day-old ferment that had been stored in the refrigerator, the second was a fresh ferment aged 2, 7 or 12 days old. The positive preference index indicates a preference for the fresh ferment. The red line shows the linear regression

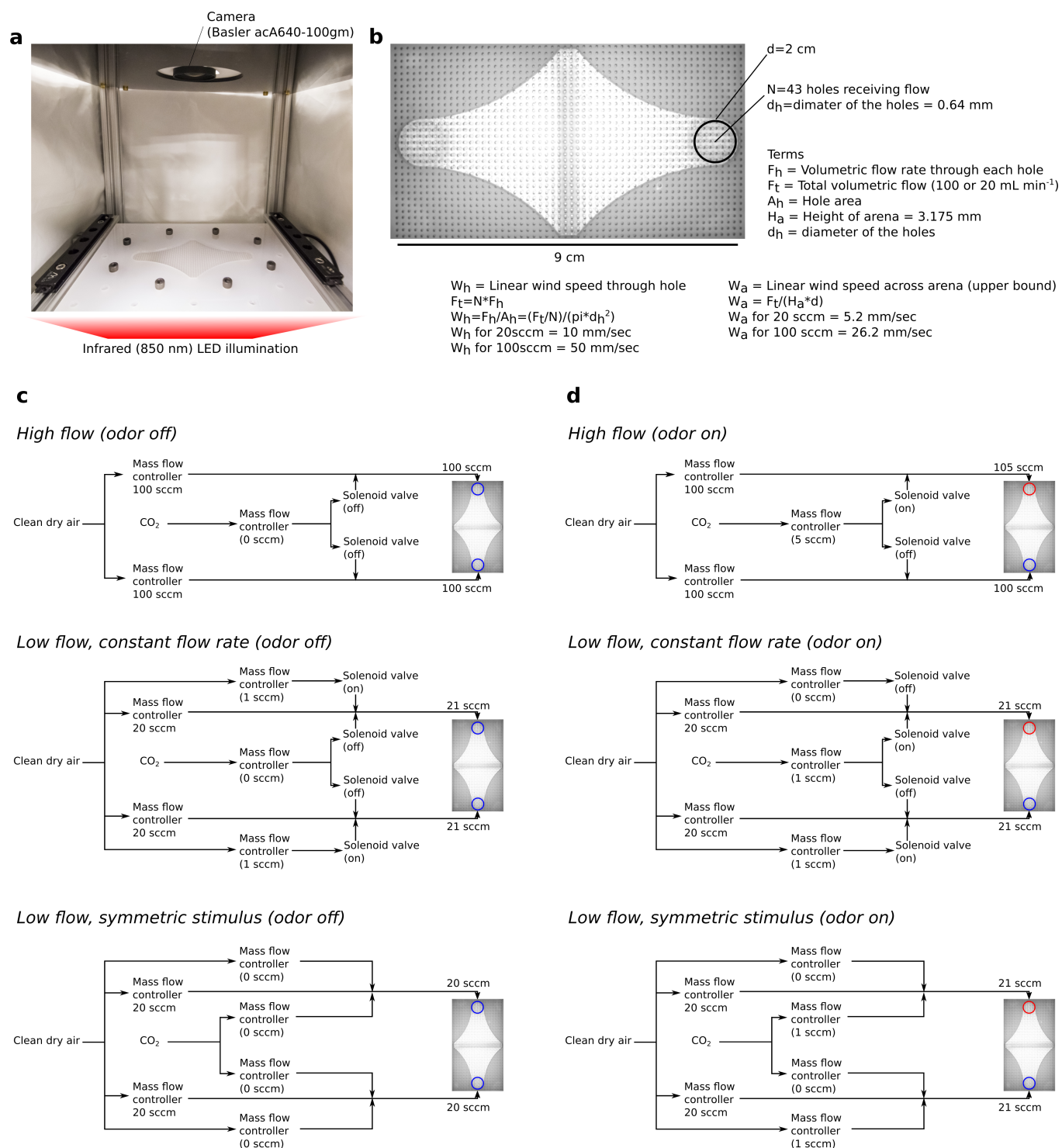
($P < 0.001$, $r^2 = 0.28$). $n = 12$ trials per condition. The mean and standard deviation of the total captured flies for each trial was 105 ± 59 . **d**, CO₂ concentration in 500-ml fly-rearing bottles under common laboratory conditions. $n = 6$ trials per condition. **e**, Measurement setup for the data shown in **d**. **f**, Time course of CO₂ concentration measurement for three bottles filled with different concentrations of CO₂. $n = 3$ per calibration gas. **g**, Peak measured CO₂ concentration versus actual CO₂ concentration for the calibration gases (black). Coloured lines show the measured peak concentrations for the actual fly-food bottles, and the resulting CO₂ concentrations shown in **d**. In all panels, shading indicates the bootstrapped 95% confidence intervals around the mean.



Extended Data Fig. 2 | See next page for caption.

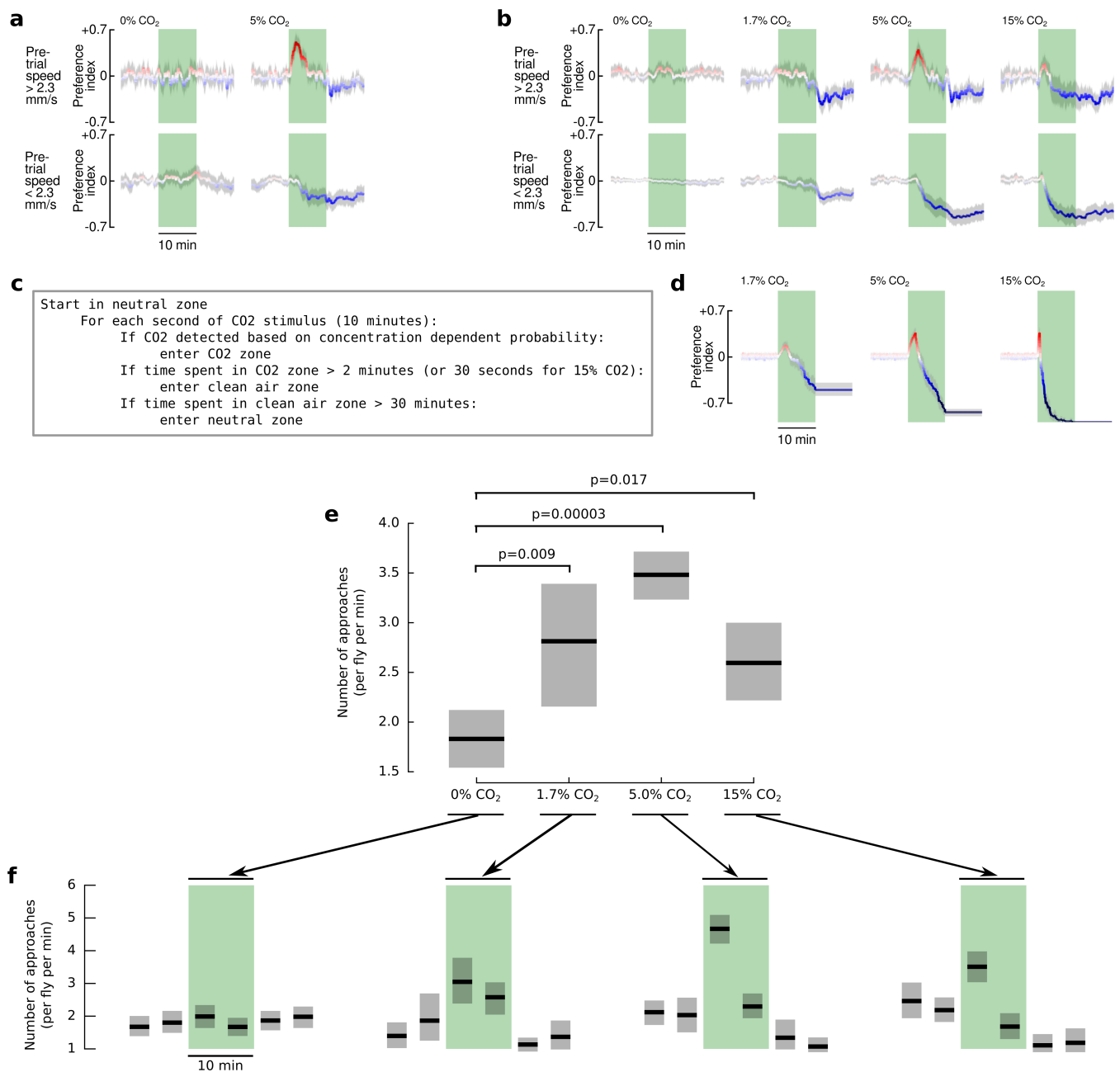
Extended Data Fig. 2 | Responses of flies to odours at different concentrations. **a**, CO₂ concentration on the landing platform (green), and at two distances downwind from the downwind edge of the platform (red and purple). Measurements (shown with points) were made for low flow rates (shown in the inset), and values at larger flowrates were extrapolated based on a linear model for measurements made at the 2-cm distance. This was necessary because the CO₂ sensor could not accurately report concentrations higher than 0.5% CO₂. **b**, Diagram that illustrates the theoretical boundary layer used to confirm our measurements (see Methods). **c**, The responses of flies to odours is consistent across a wide range of concentrations. Data plotted as in Fig. 2e, for additional flow rates. Points indicate individual data points (each trajectory contributes a single point). For each odour, we recorded the following n = number of trajectories for each of the concentrations (listed left to right). H₂O, 128, 183 and 79; CO₂, 195, 106, 125 and 48; ethanol, 173, 171 and 47; and vinegar, 219, 193 and 248. In all panels, shading indicates the bootstrapped 95% confidence intervals around the median. **d–g**, Comparison of the results from experiments with the landing platform from **c** and the

constrained walking arena used in Fig. 3. Scattergram is repeated from **c**, 60 ml min⁻¹ CO₂. To compare the data from the wind tunnel experiment to the walking arena from Fig. 3, we calculated a bootstrapped time trace. The time trace is the bootstrapped mean and 95% confidence intervals for the normalized number of flies that would have been on the platform, had all the flies landed simultaneously. The green shading is only provided for reference; the odour was never turned off in these wind tunnel experiments. **e**, Time trace from **d** overlaid on the normalized number of non-starved flies near the 5% CO₂ source during the dusk time period in the walking arena, copied from Fig. 3d. **f**, Same as **d**, but for ethanol, 60 ml min⁻¹. **g**, Time trace from **f** overlaid on the normalized number of non-starved flies near the 5% ethanol source during the dusk time period in the walking arena. Data are not shown, but are very similar to Fig. 3d ethanol case with starved flies. We chose non-starved flies for the comparisons because wind tunnel experiments were done with non-starved flies. We chose the 60 ml min⁻¹ case because the CO₂ concentration in the wind tunnel matches the 5% CO₂ stimulus in the walking experiments.



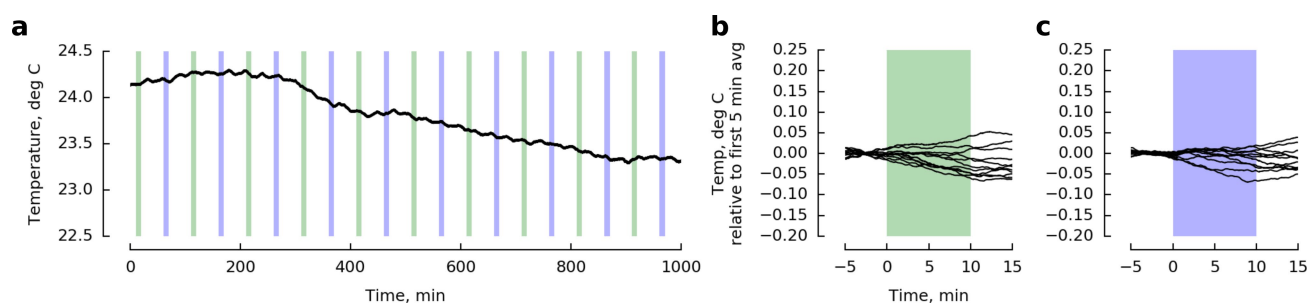
Extended Data Fig. 3 | Walking arena geometry and odour stimulus.
a, Photograph of walking arena, with the lid removed. **b**, Annotated photograph of the walking arena as seen from above, taken with the machine vision camera that is used for tracking. **c**, Odour control for

the three delivery architectures, with odour off. **d**, Odour control for the three delivery architectures, with odour on. In our experiments, the port through which odour is delivered was alternated.



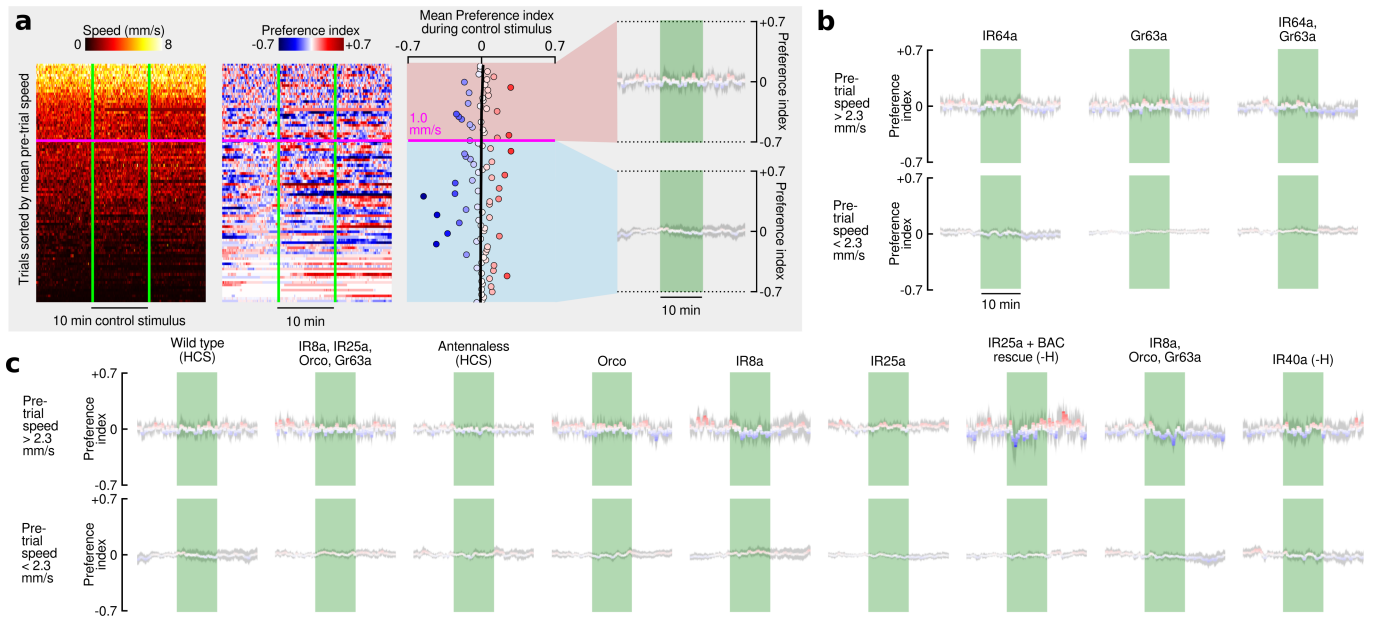
Extended Data Fig. 4 | Responses to CO₂ are strongest at 5% concentration and are unaffected by social dynamics. **a**, Control and 5% CO₂ responses for individual flies. For these experiments, we starved a single two-day old wild-type (HCS) female fly for either 24 h or 3 h before starting the experiment. In every other way, the data are plotted as in Fig. 4. The data shown were collected from $n = 29$ individual flies, in which each fly was subject to a 20-h long experiment with $n = 14$ 5% CO₂ stimuli and $n = 10$ control stimuli. **b**, CO₂ responses exhibited by flies to three concentrations of CO₂. For these experiments, we starved groups of 10 flies for 24 h before starting the experiment. Flies were presented with 0%, 1.7% or 5% CO₂ in one set of experiments, and 0% or 15% in another set. Data are plotted as in Fig. 4. $n = 20$ –170 trials per condition. To explain the complex dynamics of the approach behaviour under the different CO₂ concentrations, we made a very simple agent-based model with the pseudocode shown in **c**; see Supplementary Information for additional discussion. **d**, Dynamics of the CO₂ attraction of flies can be explained by

the simple agent-based model described in **c**. Preference indices are shown for the results of $n = 100$ iterations of the model, under three different CO₂ concentrations. The data are plotted in the same manner as **b**. The key insight offered by this model is that although our agents were programmed to exhibit the same behaviour towards 1.7% and 5% CO₂, the decreased likelihood of them detecting the lower concentration CO₂ in conjunction with the long-term aversion results in an apparent indifference towards low concentrations of CO₂. **e**, To show that flies are indeed attracted to the low (1.7%) concentration of CO₂, we used a different analysis that calculated the number of times that flies approached the CO₂ source during the course of each 10-min stimulus. Pairwise statistics were determined with the two-sample Kolmogorov–Smirnov test (test statistics were 0.57, 0.83 and 0.41 for comparisons between 0% and 1.7%, 5%, and 15%). **f**, Time course of the number of times that flies approach the CO₂ source, in 5-min intervals. In each panel, the shading shows the bootstrapped 95% confidence intervals around the mean.

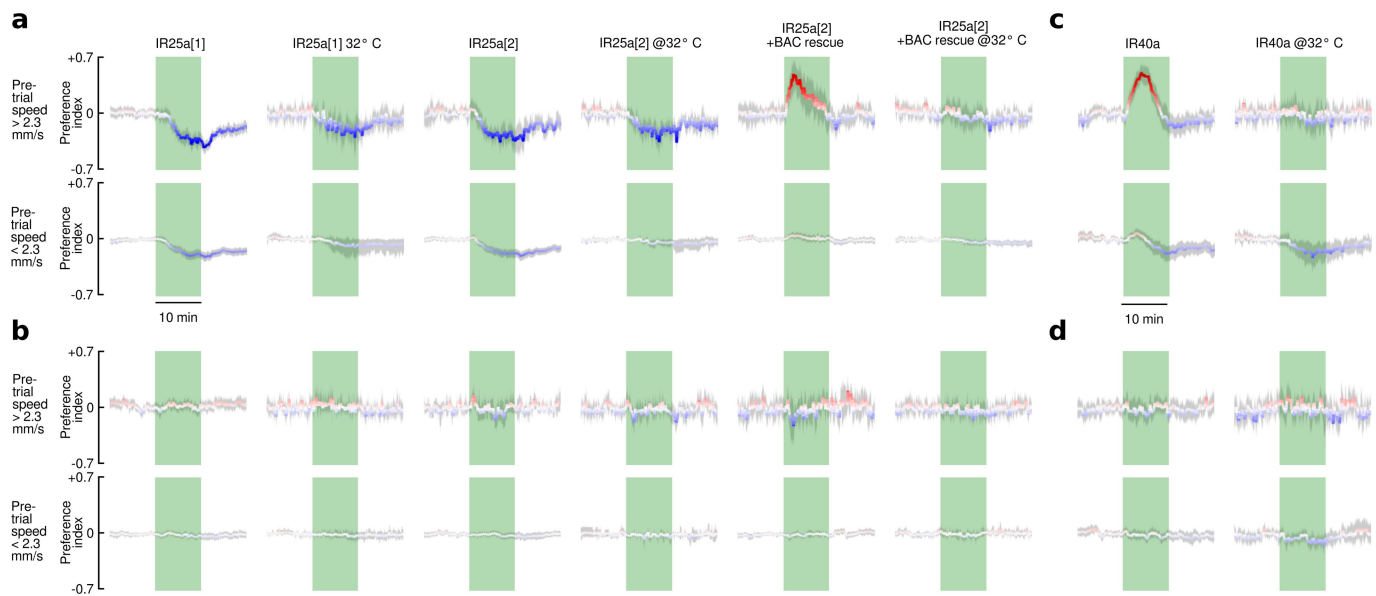


Extended Data Fig. 5 | Temperature measurements in the walking arena show no correlation with CO₂ or clean air stimuli. **a**, Temperature over the course of 16 h (see Methods). As in our experiments, every 40 min a 10-min CO₂ stimulus identical to that used in Fig. 4 was applied either to

the side of the arena with the temperature probe (green shading) or to the opposite side of the arena (blue shading). **b**, **c**, Data from a time-aligned and baseline-subtracted for CO₂ and control trials, respectively.

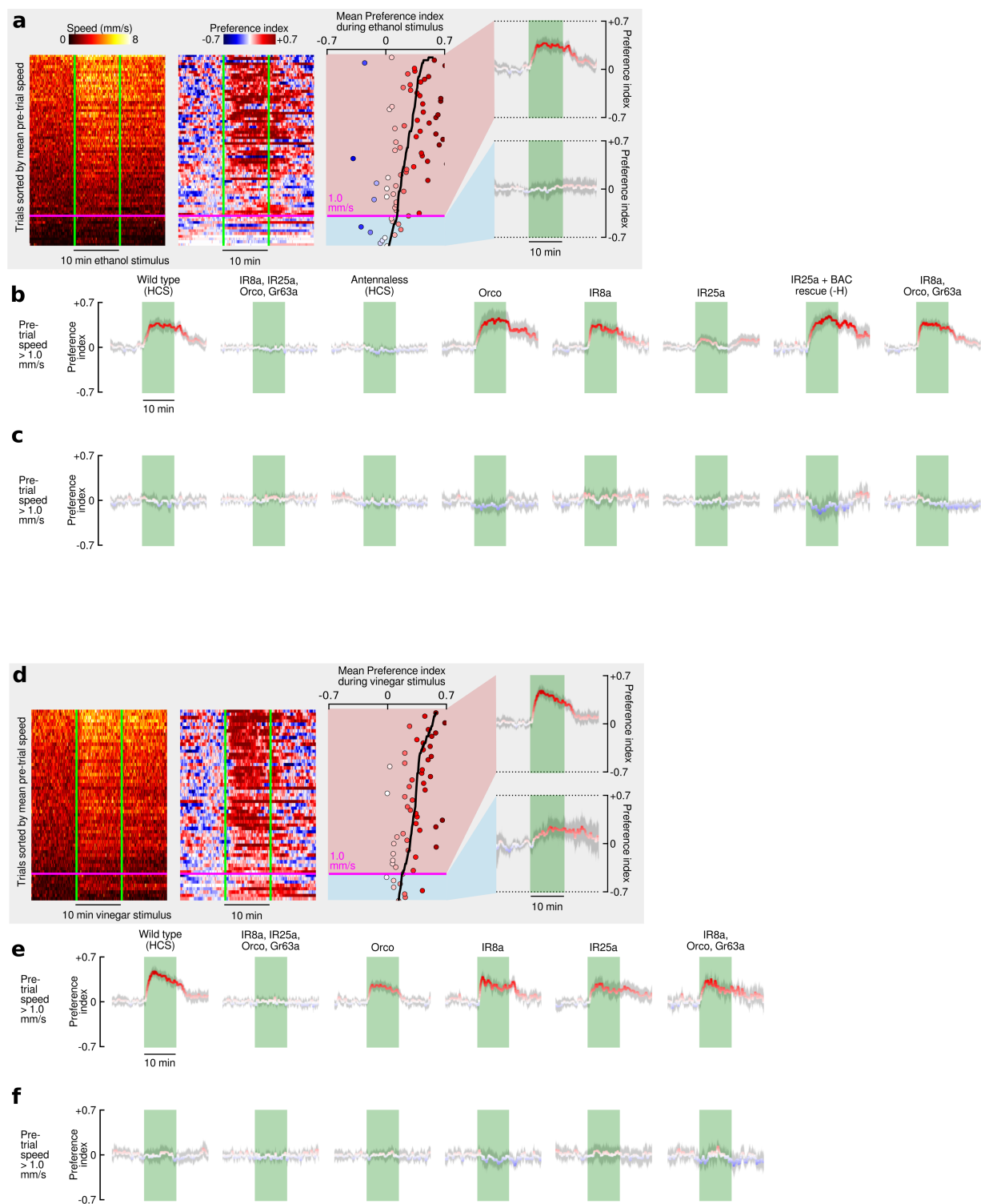


Extended Data Fig. 6 | Flies do not respond to a stimulus of clean air (without CO₂). Data plotted as in Fig. 4, but for a 0% CO₂ stimulus. $n = 17\text{--}81$ trials per condition. Shading indicates the bootstrapped 95% confidence intervals around the mean.



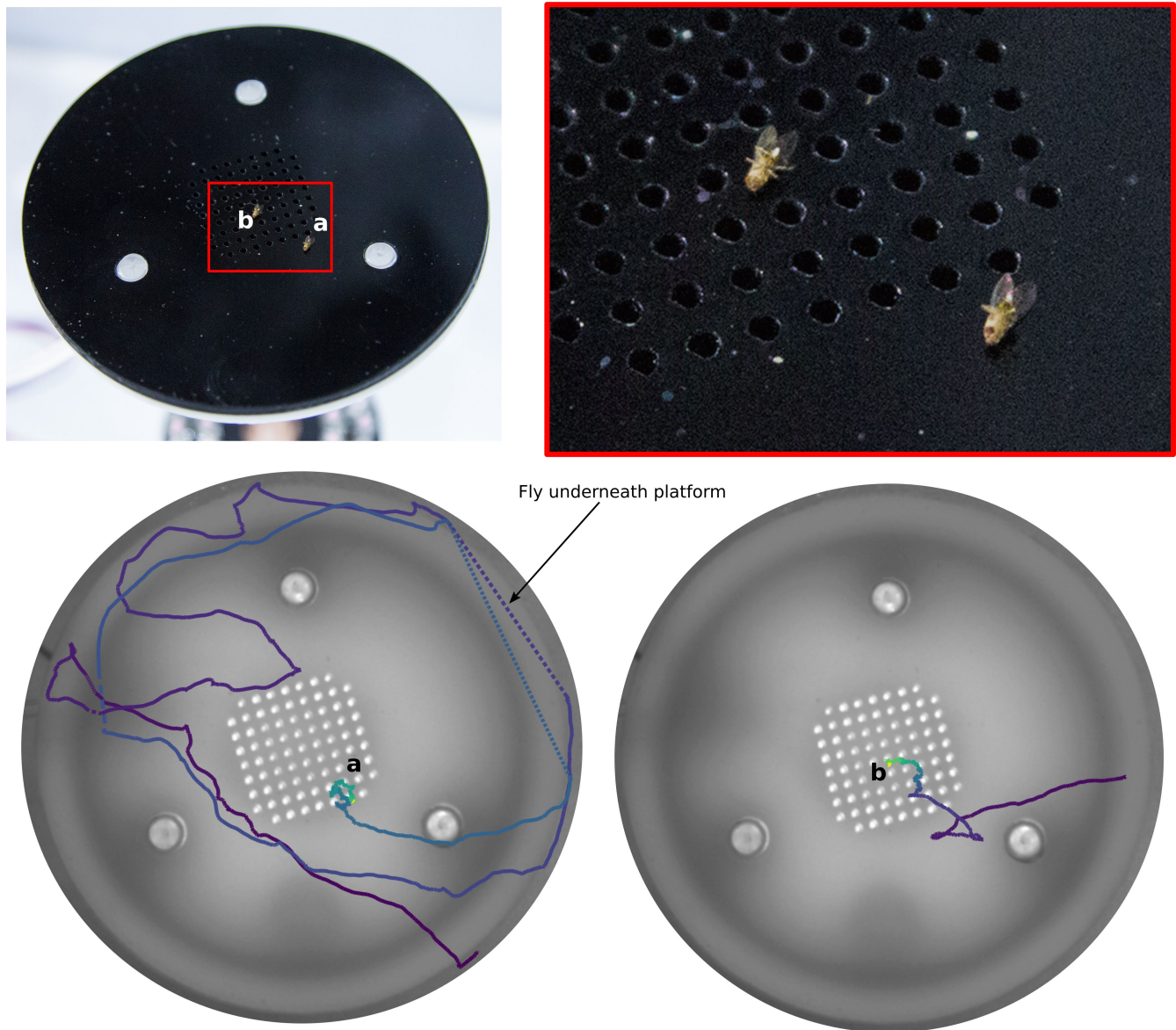
Extended Data Fig. 7 | IR25a is required for CO₂ attraction and IR40a is not. As in Fig. 4, the data from each experimental group are sorted according to the mean speed during the reference period of 5 min before the odour stimulus. In addition, for each mutant we show two sets of panels corresponding to: (1) flies that were starved for 24 h or 3 h before experiments conducted at 23 °C, and (2) flies that were starved for 3 h before experiments done at 32 °C. This arrangement is

in contrast to Fig. 4, in which data from the two temperature groups are combined. **a, b**, Responses of two IR25a mutants and a bacterial artificial chromosome rescue to a 5% CO₂ stimulus (**a**) and a 0% CO₂ stimulus (**b**). **c, d**, Responses of an IR40a mutant to a 5% CO₂ stimulus (**c**) and a 0% CO₂ stimulus (**d**). $n = 4-78$ trials per condition. Shading indicates the bootstrapped 95% confidence intervals around the mean.



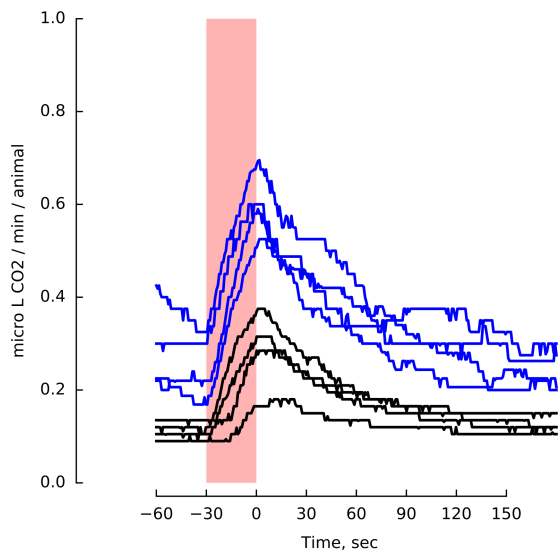
Extended Data Fig. 8 | IR25a is required for ethanol attraction but not vinegar attraction. Data plotted as in Fig. 4. Experiments were done with 24-h-starved flies only. **a, b**, Responses to 3 ml min^{-1} air passed through a bottle of pure ethanol added to 20 ml min^{-1} clean air. **c**, Control responses with 3 ml min^{-1} of clean air added to 20 ml min^{-1} of clean air.

d, e, Responses to 3 ml min^{-1} air passed through a bottle of pure vinegar added to 20 ml min^{-1} clean air. **f**, Control responses with 3 ml min^{-1} of clean air added to 20 ml min^{-1} of clean air. $n = 14-70$ trials per condition. Shading indicates the bootstrapped 95% confidence intervals around the mean.



Extended Data Fig. 9 | *Drosophila* are attracted to fatal levels of CO₂. Top, Photograph of two flies that were fatally attracted to a 200 ml min⁻¹ CO₂ stimulus. Bottom, trajectories for these two flies before they became

anaesthetized and died. Colour encodes time, starting at purple and ending at green or yellow.



Extended Data Fig. 10 | Flies and mosquitoes both increase CO₂ production when shaken. Red shading indicates the time during which the vial was shaken. We tested four groups of 10–20 animals for flies (black) and mosquitoes (blue). CO₂ was measured with a LiCorr-6262. See Methods for details.

Complex mammalian-like haematopoietic system found in a colonial chordate

Benjamin Rosental^{1,2,11*}, Mark Kowarsky^{3,11}, Jun Seit^{1,4}, Daniel M. Corey¹, Katherine J. Ishizuka^{1,2}, Karla J. Palmeri^{1,2}, Shih-Yu Chen⁵, Rahul Sinha¹, Jennifer Okamoto⁶, Gary Mantalas^{7,8}, Lucia Manni⁹, Tal Raveh¹, D. Nathaniel Clarke², Jonathan M. Tsai¹, Aaron M. Newman¹, Norma F. Neff⁶, Garry P. Nolan⁵, Stephen R. Quake^{6,7,12}, Irving L. Weissman^{1,2,10,12*} & Ayelet Voskoboynik^{1,2,12*}

Haematopoiesis is an essential process that evolved in multicellular animals. At the heart of this process are haematopoietic stem cells (HSCs), which are multipotent and self-renewing, and generate the entire repertoire of blood and immune cells throughout an animal's life¹. Although there have been comprehensive studies on self-renewal, differentiation, physiological regulation and niche occupation in vertebrate HSCs, relatively little is known about the evolutionary origin and niches of these cells. Here we describe the haematopoietic system of *Botryllus schlosseri*, a colonial tunicate that has a vasculature and circulating blood cells, and interesting stem-cell biology and immunity characteristics^{2–8}. Self-recognition between genetically compatible *B. schlosseri* colonies leads to the formation of natural parabionts with shared circulation, whereas incompatible colonies reject each other^{3,4,7}. Using flow cytometry, whole-transcriptome sequencing of defined cell populations and diverse functional assays, we identify HSCs, progenitors, immune effector cells and an HSC niche, and demonstrate that self-recognition inhibits allospecific cytotoxic reactions. Our results show that HSC and myeloid lineage immune cells emerged in a common ancestor of tunicates and vertebrates, and also suggest that haematopoietic bone marrow and the *B. schlosseri* endostyle niche evolved from a common origin.

Charles Darwin recognized that the study of tunicates is critical to understand the evolution of vertebrates, and tunicates were later discovered to be a sister group of vertebrates^{9–11}. To gain insight into the evolution of the mammalian haematopoietic system, we characterized the haematopoietic and immune system in the colonial tunicate *B. schlosseri*.

B. schlosseri colonies produce genetically identical individuals (zooids) through stem-cell-mediated cyclical budding⁵ (Fig. 1a, b). Every week, developed buds replace their parent zooids, which then undergo synchronized programmed cell death¹² (Supplementary Video 1). When colonies touch, their extracorporeal vasculatures either fuse or reject^{2,3} (Fig. 1c, Supplementary Video 2). This self–nonself recognition process is controlled by the highly polymorphic histocompatibility gene *BHF*, and at least one shared *BHF* allele is required for fusion to take place⁷. We adapted fluorescence-activated cell sorting¹³ (FACS) to separate *B. schlosseri* cells and isolated 34 cell populations using size, granularity, natural auto-fluorescence, reagents such as antibodies that bind differentially to live cells (CD49d, CD57 and BHF), concanavalin-A, and alkaline phosphatase expression (Extended Data Figs. 1, 2, Extended Data Tables 1, 2a). We sequenced the transcriptomes of 23 sorted cell populations, the hierarchical endpoint populations of our FACS gating strategy (Extended Data Fig. 1c, d, Supplementary

Table 1), and found correlations between gene expression profiles, morphology and marker expression (Extended Data Fig. 3). In the cluster of cell populations CP25, CP33 and CP34, there were 235 genes that were differentially upregulated and are known to be expressed in vertebrate blood and haematopoietic systems¹⁴ (Extended Data Fig. 4a, Supplementary Table 2). Analysis of this gene set by Gene Expression Commons¹⁵ against gene expression data from 39 distinct mouse haematopoietic stem, progenitor and differentiated cells revealed significant overlap ($P < 0.05$) in expression between CP25, CP33 and CP34, and mammalian haematopoietic stem, progenitor and myeloid lineage cells (Fig. 2a, Supplementary Tables 2, 3).

To measure the ability of these candidate HSC populations (cHSCs) to differentiate into other cell types, cells were transplanted from orange

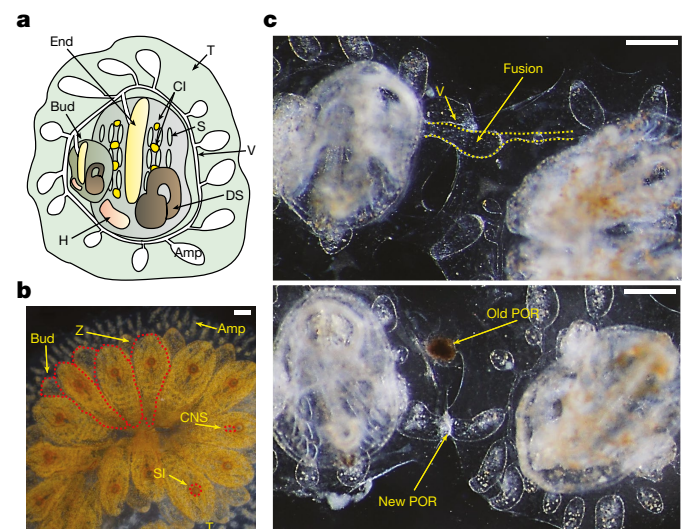


Fig. 1 | *B. schlosseri* anatomy and natural transplantation reactions. **a**, Diagram of a zooid (ventral view) and primary bud (Bud), embedded within a tunic (T), with vasculature (V) connected to the zooid and bud, which terminates in ampullae (Amp). The zooid has a branchial sac consisting of the endostyle (End) and stigmata (S), cell islands (CI), digestive system (DS) and heart (H). **b**, Live imaging of a colony (dorsal view). Developing buds are connected to the parental zooids (Z); all are connected to blood vessels. The zooid's siphons (SI) and central nervous system (CNS) are observed. **c**, Live imaging of colonies undergoing fusion (top) and rejection (bottom). Arrows point to fused vasculature and points of rejection (PORs). Scale bar, 0.2 mm.

¹Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, CA, USA. ²Department of Biology, Stanford University, Hopkins Marine Station, Pacific Grove, CA, USA. ³Department of Physics, Stanford University, Stanford, CA, USA. ⁴AI based Healthcare and Medical Data Analysis Standardization Unit, Medical Sciences Innovation Hub Program, RIKEN, Tokyo, Japan. ⁵Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA, USA. ⁶Chan Zuckerberg Biohub, San Francisco, CA, USA. ⁷Department of Bioengineering, Stanford University, Stanford, CA, USA. ⁸Department of Molecular Cellular and Developmental Biology, University of California Santa Cruz, Santa Cruz, CA, USA. ⁹Dipartimento di Biologia, Università degli Studi di Padova, Padova, Italy. ¹⁰Ludwig Center for Cancer Stem Cell Research and Medicine, Stanford University School of Medicine, Stanford, CA, USA. ¹¹These authors contributed equally: Benjamin Rosental, Mark Kowarsky. ¹²These authors jointly supervised this work: Stephen R. Quake, Irving L. Weissman, Ayelet Voskoboynik. *e-mail: rosentab@post.bgu.ac.il; irv@stanford.edu; ayeletv@stanford.edu

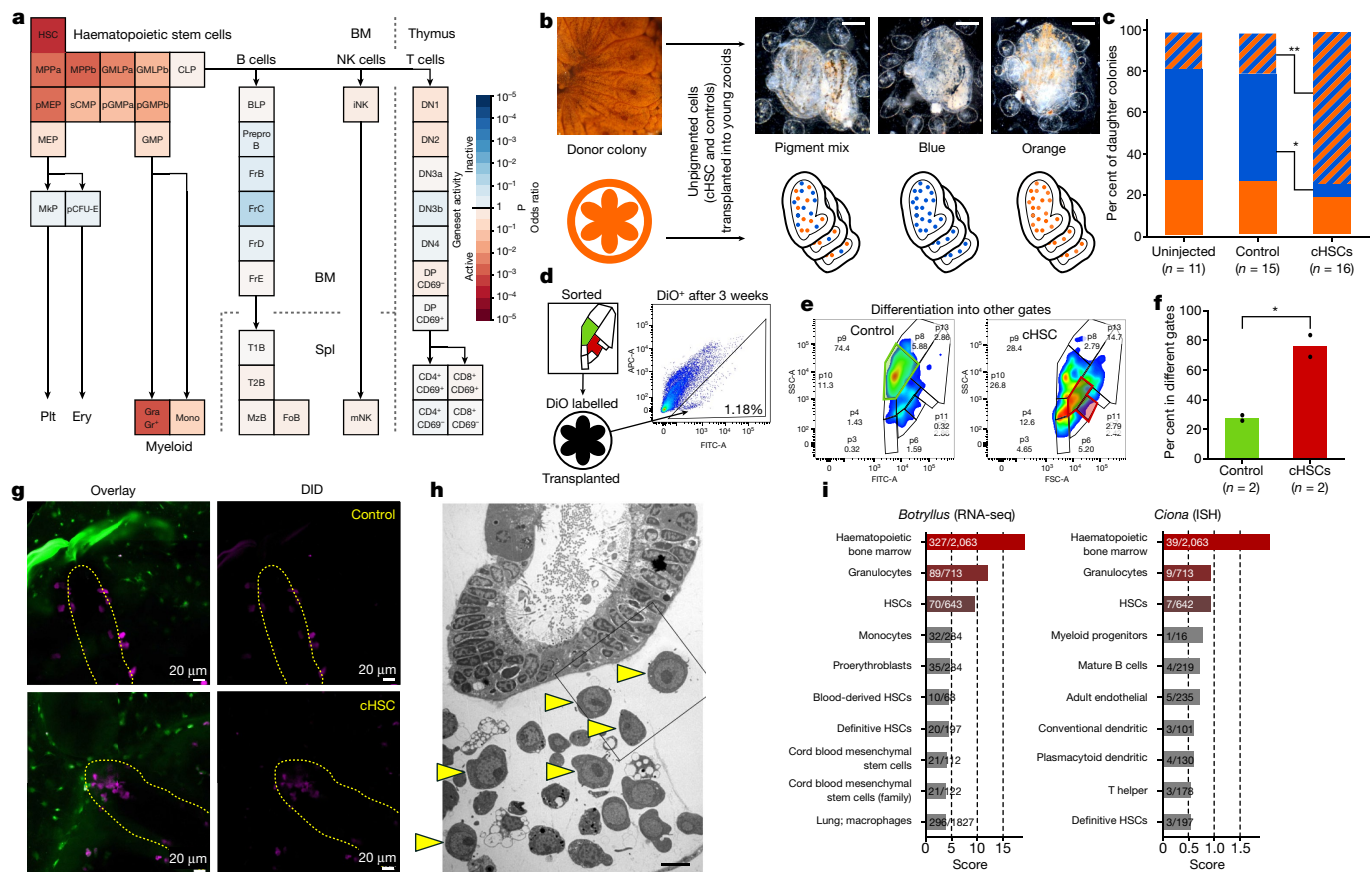


Fig. 2 | Multilineage differentiation capacity, homing sites of cHSCs and their niches. **a**, Geneset Activity Analysis genes upregulated ($n = 235$) in candidate HSCs (CP25, CP33 and CP34) using the Gene Expression Commons tool on a mouse haematopoiesis model; each box is a defined cell population in the haematopoietic process. The enriched populations are HSCs and the myeloid lineage. **b**, Candidate HSCs and a control cell population (CP18) from an orange-pigmented donor colony were transplanted into compatible recipient colonies with blue, orange or a mixture of the two pigmented cells. Top, live imaging. Scale bar, 0.2 mm. Bottom, schematics. **c**, Significant reduction in the ratio of blue colonies and significant increase in the ratio of mixed pigmented colonies (Fisher's exact test, two-tailed, $*P = 0.006$, $**P = 0.004$), 20 days after cHSC transplantation. **d–f**, cHSC and a control cell population (CP18) were labelled with DiO and transplanted into compatible colonies (**d**). Three weeks after transplantation, DiO⁺ cells (**d**, right) from the recipient colonies were analysed by FACS. Twenty-four per cent of the cHSC transplanted cells were detected in their original gate (**e**, right, in red) and the rest were detected in other gates. The majority of the transplanted cells

donor colonies to compatible colonies (Fig. 2b). Twenty days after transplantation, the unpigmented cHSC populations (CP25, CP33 and CP34) from the orange donor colony differentiated into orange pigmented cells within the recipient colonies, as shown by upregulation of mixed pigmented colonies and a reduction in pure blue colonies, a trend that was not observed in control (CP18) or uninjected colonies (Fig. 2c). To test the capacity of cHSCs to undergo multilineage differentiation, the cHSC population and control population were isolated, labelled with a fluorescent membrane dye (DiO), and transplanted into compatible recipient colonies. Three weeks after transplantation, 76% of transplanted cHSCs were detected in a different FACS gate than the original, as compared with less than 30% of transplanted cells from the control group (Fig. 2d–f, Extended Data Fig. 5a).

To identify the cHSC niches, the cHSC population and a control population (CP3) were isolated, labelled with a lipophilic dye (DiD) and injected into compatible colonies labelled with carboxyfluorescein succinimidyl ester (CFSE). Five to ten days after transplantation, DiD-labelled cHSC populations migrated into the recipient colony and

from the control group were detected in the original gate (**e**, left, in green). The experiment was performed twice. **f**, Percentage of cells from cHSC and control populations detected in gates different from their original gate, three weeks after transplantation. The experiment was performed on two pools of five animals of each population; unpaired *t*-test, two-tailed, $*P = 0.024$, bars show mean. **g**, cHSC and a control population (CP3) were isolated, labelled with DiD (violet) and transplanted into CFSE (green)-labelled compatible colonies. Five to ten days after transplantation, only the cHSC populations homed to the endo-niche (5/6) compared to control (0/4). $P = 0.048$; Fisher's exact test, two-tailed. Yellow dashed outline indicates endostyle; scale bar, 20 μ m. **h**, Electron microscopy section of the endostyle and subendostylar sinus (endo-niche). Yellow arrowheads indicate cells with haemoblast (cHSC) morphology. Performed three times. Scale bar, 5 μ m. **i**, Enriched blood-associated mammalian cell types for highly expressed genes in the *B. schlosseri* endostyle (RNA sequencing (RNA-seq)) and the solitary tunicate *C. robusta* endostyle (in situ hybridization)²¹.

aggregated in two known *B. schlosseri* stem-cell niches: the endostyle and the cell islands. The endostyle niche, located at the anterior subendostylar sinus (endo-niche), has been identified as a somatic stem-cell niche^{6,16} (Fig. 2g, Extended Data Fig. 5b). Proliferating cells and haemoblasts have been identified near the endostyle and around the branchial sac's stigmata in juvenile *Ciona intestinalis*¹⁷. Indeed, blood cells with a haemoblast (cHSC) morphology and proliferating cells are abundant in the *B. schlosseri* endo-niche⁶ (Fig. 2h, Extended Data Fig. 5c, d). The control population (CP3) does not express a gene signature of the haematopoietic system, but does have a germline gene expression signature and is localized to a known germline stem-cell niche¹⁶, the cell islands (Extended Data Figs. 4c, 5b). The *B. schlosseri* cHSCs localized towards the endo-niche, similar to the homing process of mammalian HSCs to the bone marrow^{18–20}. The endostyle is a complex tissue with defined anatomical structures and molecular features^{6,16,21–24}. We sequenced the transcriptomes of ten endostyles and compared them to the transcriptomes of whole colonies ($n = 34$). Homologous genes that were

significantly upregulated ($P < 0.05$) in the endostyle were analysed by GeneAnalytics¹⁴, revealing shared expression of 327 genes between the endostyle and human haematopoietic bone marrow (Fig. 2i, left, Extended Data Fig. 6a, Supplementary Tables 2, 4). This finding was further supported by Gene Expression Commons^{15,25} (Extended Data Fig. 6b). We queried previously obtained *Ciona robusta* in situ expression data²¹ for endostyle-associated genes and found them to be similar (Fig. 2i, right, Extended Data Fig. 6c).

In mammals, innate cellular immune responses are mediated in part by phagocytosis (the engulfment of target cells), and both adaptive and innate immune responses in part by cellular cytotoxicity (the direct killing of target cells). We used diverse ex vivo phagocytosis assays to identify phagocytic cells and track their cell populations (Extended Data Fig. 7). These assays revealed three major phagocytic populations, including two previously described phagocytic cells: amoebocytes (within CP4 and CP18), large phagocytes^{26,27} (within CP13) and a previously undescribed population, the candidate myeloid cell population (within CP7 and CP10) (Extended Data Fig. 7). The myeloid cells were the main contributors to phagocytosis (they contributed more than 40% to each of the phagocytosis assays), whereas the large phagocytes contributed mainly to allogeneic phagocytosis (Extended Data Fig. 7a).

We did not find a cell population with a clear mammalian cytotoxic gene expression signature. Morula cells, which contain phenoloxidase, accumulate at rejection points and have been proposed to be cytotoxic cells that mediate rejection^{26,28,29}. We detected morula cells in population CP18, but in lower levels than expected, prompting us to look for a candidate precursor cell. We studied the large granular lymphocyte-like (LGL) cells (enriched in CP31) as a potential candidate, because their morphology resembles that of natural killer cells, which mediate cytotoxicity as part of the mammalian innate immune response³⁰. Ex vivo experiments revealed that purified LGL cells transitioned into morula cells after two days (Fig. 3a, labelled 1–4), suggesting that LGL cells are cytotoxic morula cells that become pigmented granular morula cells following activation.

Assays that compared the ability of morula and LGL cells to induce cytotoxicity showed that isolated LGL cells were significantly ($P < 0.005$) more cytotoxic than isolated morula cells or the control population (CP3; Extended Data Fig. 8a). Overnight incubations of isolated LGL cells in either a syngeneic (self) or an allogeneic (nonself) challenge led to a transition of 60% of the LGL cells that were incubated with allogeneic cells and an increase in morula cells. Only about 10% of the LGL cells that were incubated with syngeneic cells became morula cells (Extended Data Fig. 8b). Upon activation, LGL cells change their morphology, develop granularity and pigmentation, presumably owing to phenoloxidase activation, and become morula cells (Fig. 3a, Extended Data Fig. 8b, c). This set of experiments demonstrates that the LGL cells are the cytotoxic cell population, and therefore we call LGL cells cytotoxic morula cells from here onwards.

Analysis of the genes that were differentially upregulated by the highly enriched cytotoxic morula cell population (CP31) revealed 52 unannotated genes with no human or mouse homologues. Twenty-one of these genes carry domains that are associated with functions of cytotoxic cells such as recognition or lysis, and C-type lectin, a domain contained in human natural killer cell receptors (Supplementary Table 5). Among the 18 differentially upregulated genes that have sequence homology to genes in vertebrates, 14 are associated with at least one of the following functions: cellular recognition, cytotoxicity or peptidase activity, leukocyte homing and general immune response (Fig. 3b). CP31 cells also express tyrosinase-associated gene (*TYRP2*), the vertebrate paralogue gene to phenoloxidase, which is one of the main enzymes found in morula cells²⁸ (present at sevenfold-higher levels than in other cells). Although we did not find a *B. schlosseri* cell population that had a significant lymphoid lineage signature, Geneset Activity Analysis suggested that population CP19—which is enriched in morula cells—resembles mouse T cells, B cells, and immature natural killer cells in gene expression (Extended Data Fig. 4d). This list does

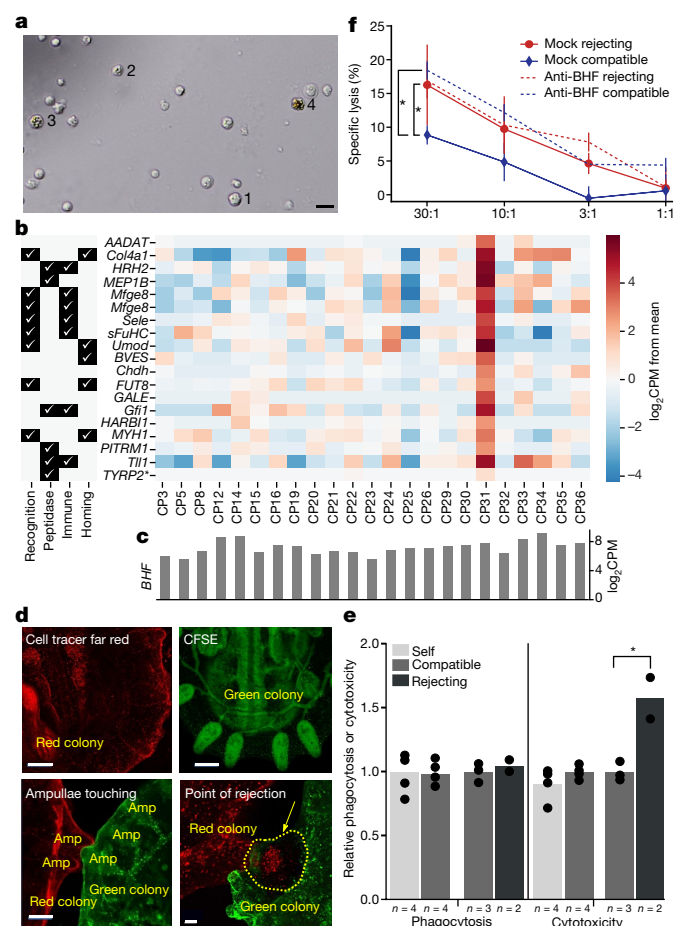


Fig. 3 | Allorecognition is mediated by cytotoxic cells through BHF recognition. **a**, Observation of isolated large granular lymphocyte-like (LGL) cells after two days ex vivo. 1, morphology of original isolated cells; 2, 3, granular and light pigmented cells; 4, granular pigmented morula cell. The experiment was performed three times. **b**, Representation of homologue annotated genes that are differentially upregulated by LGL-enriched population CP31. *Mfge8* is listed twice because two upregulated *B. schlosseri* genes were homologues of this mouse gene. **TYRP2* (paralogue of phenoloxidase, also known as *Dct*) is not differentially expressed, but is sevenfold higher in CP31. Left side is a table indicating gene association with immunity functions. CPM, counts per million. **c**, BHF expression levels in the cell populations. **d**, Colony labelled with cell tracker far red (top left), and with CFSE in green (top right). Bottom left, ampullae touching; bottom right, points of rejection; a mixture of allogeneic cells from the two colonies is observed in the area of necrotic tissue (yellow dashed line). The experiment was repeated three times. Scale bar, 0.2 mm. **e**, Phagocytosis and cytotoxicity assays were set between compatible, incompatible or self. Significant upregulation of cytotoxicity between cells taken from incompatible colonies is observed. Experiment used $10^5:10^5$ cell ratio and were performed twice with 2–4 samples in each treatment, and validated by additional experiments with different ratios. Unpaired *t*-test, two-tailed; * $P = 0.023$, bars show mean. **f**, Blocking of BHF with anti-BHF or mock serum. Blue, compatible colony target cells; red, incompatible colony target cells. The blocking of BHF significantly upregulated the cytotoxicity of compatible targets compared to mock serum. The experiment was performed three times with duplicates or triplicates. Two-factor ANOVA with replication, * $P = 0.0008$; mean \pm s.d.

not include any gene that is associated with known adaptive immunity function (Supplementary Table 2).

To test whether allogeneic cells interact in vivo during rejection, whole colonies were differentially labelled, set near each other and monitored by live imaging (Fig. 3d, Supplementary Video 3). We detected direct contact between allogeneic cells during rejection within the points of rejection (PORs; Fig. 3d, lower right). Next, we performed allogeneic phagocytosis and cytotoxicity assays. Assays

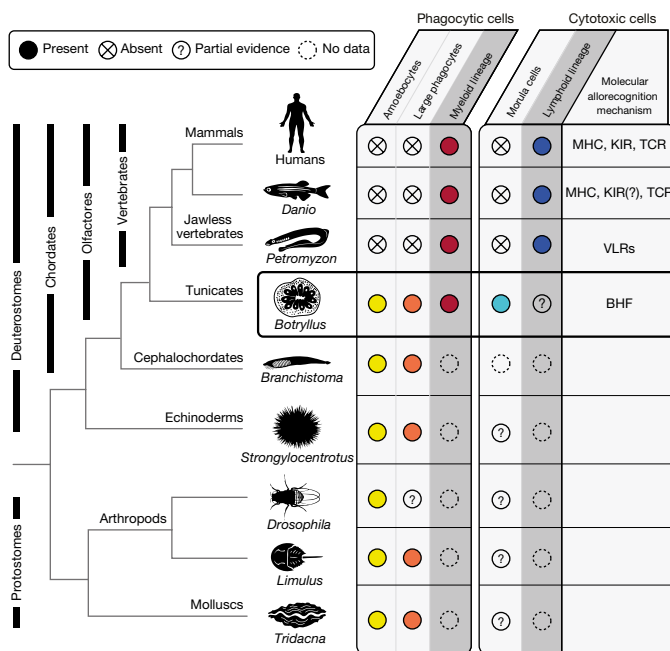


Fig. 4 | Evolution of the cellular immune system. Information regarding the cellular immune systems of invertebrate and vertebrate species. The table to the right shows the types of immune-associated cell found in each animal. Colonial tunicates contain immune system cells found in both invertebrate and vertebrate species. Whereas amoebocytes and large phagocytes are found in *B. schlosseri* and other invertebrate species (yellow and orange), myeloid lineage phagocytic cells have so far been found only in *B. schlosseri* and vertebrates (red). Cytotoxic morula cells have been identified in *B. schlosseri* and are likely to exist in a few other invertebrate species (turquoise), but have not been identified in vertebrates. Although a classic lymphoid lineage has been found only in vertebrates (blue), there is a cellular population and additional molecular analysis suggesting the existence of mainly undifferentiated lymphoid cells in *B. schlosseri*. The cellular allorecognition molecules identified in each species are shown in the right-hand column: immunoglobulin superfamily major histocompatibility complex (MHC), killer inhibitory receptors (KIRs) and T cell receptors (TCRs), leucine-rich repeats receptors of the variable lymphocyte receptors (VLRs), and BHF. Question marks represent missing functional or molecular validation. Data source summarized in Extended Data Table 2b.

were set between cells taken from compatible colonies (sharing one BHF allele), rejecting colonies (no shared BHF allele) or self (same colony) (Fig. 3e). We used confocal microscopy to validate phagocytosis; in the cytotoxicity assays, we validated specific lysis by increasing ratios between effector and target cells (Extended Data Fig. 8d, e). The experiment was accomplished by differential labelling of cells that originated from the same colony, two different compatible colonies or two rejecting colonies. There was significantly more cellular cytotoxicity ($P < 0.05$) between cells that were derived from rejecting colonies than between cells derived from either compatible colonies or the same colony (Fig. 3e, Extended Data Fig. 8e). As all the treatments led to similar levels of phagocytosis, we concluded that cytotoxicity is the immune effector mechanism of the cellular allogeneic response.

In a third set of experiments, we tested whether blocking BHF would affect the cytotoxic reaction. Cytotoxicity assays between compatible and rejecting colonies treated with mock serum revealed significantly more cytotoxicity between rejecting colonies (Fig. 3f). When we blocked BHF with serum containing polyclonal anti-BHF antibodies that bind live *B. schlosseri* cells (Extended Data Fig. 2e, Extended Data Table 2a), cell lysis between compatible colonies was enhanced to the level observed in rejecting colonies (Fig. 3f). This treatment did not affect the level of cell lysis observed in rejecting colonies (Fig. 3f). These results show that BHF is a major histocompatibility factor that is essential for self-recognition. Similar to inhibition of natural killer

cells by the major histocompatibility complex, self-BHF recognition inhibits cytotoxicity. These results—and the observation that fusion occurs when colonies share at least one BHF allele⁷—demonstrate that, as in natural killer cell recognition, the cellular cytotoxicity mechanism in *B. schlosseri* is based on ‘missing self’³¹.

The haematopoietic and immune systems of *B. schlosseri* combine features of vertebrates and invertebrates (Fig. 4). *B. schlosseri* has cells that share morphological and molecular characteristics with the vertebrate HSC and myeloid lineages, including cells that take part in phagocytosis. It also has amoebocytes and large phagocytes with morphologies resembling invertebrate cell types^{32,33}. The cytotoxic morula cells of *B. schlosseri* carry imprints reminiscent of vertebrate lymphocytes, but mainly express tunicate-specific genes (Fig. 3b, Extended Data Fig. 4d, Supplementary Tables 2, 5). Several studies have described morula-like cells with phenoloxidase activity in other invertebrate species^{33,34}. Our analysis of gene sets that are expressed by the cytotoxic morula cells and identification of the BHF inhibition pathway will be likely to reveal novel mechanisms to delimit self from non-self and target pathogens.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0783-x>.

Received: 1 December 2017; Accepted: 15 October 2018;

Published online 5 December 2018.

- Weissman, I. L. Stem cells: units of development, units of regeneration, and units in evolution. *Cell* **100**, 157–168 (2000).
- Sabbadin, A. Le basi genetiche della capacità di fusione fra colonie in *Botryllus schlosseri* (Ascidacea). *Rend. Accad. Naz. Lincei. Ser. B*, 1031–1035 (1962).
- Scofield, V. L., Schlumpberger, J. M., West, L. A. & Weissman, I. L. Protochordate allorecognition is controlled by a MHC-like gene system. *Nature* **295**, 499–502 (1982).
- Stoner, D. S., Rinkevich, B. & Weissman, I. L. Heritable germ and somatic cell lineage competitions in chimeric colonial protochordates. *Proc. Natl Acad. Sci. USA* **96**, 9148–9153 (1999).
- Laird, D. J., De Tomaso, A. W. & Weissman, I. L. Stem cells are units of natural selection in a colonial ascidian. *Cell* **123**, 1351–1360 (2005).
- Voskoboinik, A. et al. Identification of the endostyle as a stem cell niche in a colonial chordate. *Cell Stem Cell* **3**, 456–464 (2008).
- Voskoboinik, A. et al. Identification of a colonial chordate histocompatibility gene. *Science* **341**, 384–387 (2013).
- Corey, D. M. et al. Developmental cell death programs license cytotoxic cells to eliminate histocompatible partners. *Proc. Natl Acad. Sci. USA* **113**, 6520–6525 (2016).
- Darwin, C. *On the Origin of the Species by Natural Selection* (John Murray, London, 1859).
- Delsuc, F., Brinkmann, H., Chourrout, D. & Philippe, H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**, 965–968 (2006).
- Voskoboinik, A. et al. The genome sequence of the colonial chordate, *Botryllus schlosseri*. *eLife* **2**, e00569 (2013).
- Lauzon, R. J., Patton, C. W. & Weissman, I. L. A morphological and immunohistochemical study of programmed cell death in *Botryllus schlosseri* (Tunicata, Ascidacea). *Cell Tissue Res.* **272**, 115–127 (1993).
- Hulett, H. R., Bonner, W. A., Barrett, J. & Herzenberg, L. A. Cell sorting: automated separation of mammalian cells as a function of intracellular fluorescence. *Science* **166**, 747–749 (1969).
- Ben-Ari Fuchs, S. et al. GeneAnalytics: an integrative gene set analysis tool for next generation sequencing, RNA-seq and microarray data. *OMICS* **20**, 139–151 (2016).
- Seita, J. et al. Gene Expression Commons: an open platform for absolute gene expression profiling. *PLoS ONE* **7**, e40321 (2012).
- Rinkevich, Y. et al. Repeated, long-term cycling of putative stem cells between niches in a basal chordate. *Dev. Cell* **24**, 76–88 (2013).
- Ermak, T. H. in *Phylogeny of Thymus and Bone Marrow-Bursa Cells* 45–56 (Elsevier, 1976).
- Adams, G. B. et al. Stem cell engraftment at the endosteal niche is specified by the calcium-sensing receptor. *Nature* **439**, 599–603 (2006).
- Yusuf, R. Z. & Scadden, D. T. Homing of hematopoietic cells to the bone marrow. *J. Vis. Exp.* **25**, 1104 (2009).
- Wright, D. E., Wagers, A. J., Gulati, A. P., Johnson, F. L. & Weissman, I. L. Physiological migration of hematopoietic stem and progenitor cells. *Science* **294**, 1933–1936 (2001).
- Ogasawara, M. et al. Gene expression profiles in young adult *Ciona intestinalis*. *Dev. Genes Evol.* **212**, 173–185 (2002).

22. Cañestro, C., Bassham, S. & Postlethwait, J. H. Evolution of the thyroid: anterior-posterior regionalization of the *Oikopleura* endostyle revealed by *Otx*, *Pax2/5/8*, and *Hox1* expression. *Dev. Dyn.* **237**, 1490–1499 (2008).
23. Burighel, P. & Cloney, R. A. in *Microscopic Anatomy of Invertebrates* (eds Harrison, F. W. & Ruppert, E. E.) 221–347 (Alan R. Liss, 1997).
24. Ogasawara, M., Lauro, R. D. & Satoh, N. Ascidian homologs of mammalian thyroid transcription factor-1 gene are expressed in the endostyle. *Zool. Sci.* **16**, 559–565 (1999).
25. Chen, J. Y. et al. *Hoxb5* marks long-term haematopoietic stem cells and reveals a homogenous perivascular niche. *Nature* **530**, 223–227 (2016).
26. Ballarin, L. & Cima, F. Cytochemical properties of *Botryllus schlosseri* haemocytes: indications for morpho-functional characterisation. *Eur. J. Histochem.* **49**, 255–264 (2005).
27. Lauzon, R. J., Brown, C., Kerr, L. & Tiozzo, S. Phagocyte dynamics in a highly regenerative urochordate: insights into development and host defense. *Dev. Biol.* **374**, 357–373 (2013).
28. Franchi, N. & Ballarin, L. Immunity in protochordates: the tunicate perspective. *Front. Immunol.* **8**, 674 (2017).
29. Oren, M. et al. Marine invertebrates cross phyla comparisons reveal highly conserved immune machinery. *Immunobiology* **218**, 484–495 (2013).
30. Timonen, T., Ortaldo, J. R. & Herberman, R. B. Characteristics of human large granular lymphocytes and relationship to natural killer and K cells. *J. Exp. Med.* **153**, 569–582 (1981).
31. Ljunggren, H. G. & Kärre, K. In search of the ‘missing self’: MHC molecules and NK cell recognition. *Immunol. Today* **11**, 237–244 (1990).
32. Smith, L. C. et al. in *Invertebrate Immunity* (ed. Söderhäll, K.) 260–301 (Springer, 2010).
33. Kawabata, S. in *Invertebrate Immunity* (ed. Söderhäll, K.) 122–136 (Springer, 2010).
34. Ballarin, L. Ascidian cytotoxic cells: state of the art and research perspectives. *Inv. Surv. J.* **9**, 1–6 (2012).

Acknowledgements We thank C. Lowe, C. Anselmi, I. Dimov, S. Karten, C. Patton, J. Thompson, P. Lovelace, R. Voskoboynik, N. Fernhoff, W.-J. Lu, P. Chu, K. Weiskopf, M. Oren, B. Wang, J. Lee, B. Compton, K. Uhlinger, T. Naik

and T. Storm for technical advice and help. This study was supported by NIH grants R56AI089968, R01AG037968 and R01GM100315 (to I.L.W., S.R.Q., and A.V.), the Virginia and D. K. Ludwig Fund for Cancer Research, a grant from the Siebel Stem Cell Institute and a Stinehart-Reed grant (to I.L.W.). L.M. was supported by PRIN - Prot. 2015NSFHXF. B.R. was supported by a Postdoctoral Fellowship of the Human Frontier Science Program Organization LT000591/2014-L, NIH Immunology training grant 5T32AI07290-28 and NIH Hematology training grant T32 HL120824-03.

Reviewer information *Nature* thanks M. D. Cooper, W. Jeffery, G. Litman and J. Pascual-Anaya for their contribution to the peer review of this work.

Author contributions Conception and design: B.R., A.V., M.K., S.R.Q. and I.L.W.; mariculture: K.J.I. and K.J.P.; flow cytometry and sorting: B.R.; CyTOF screening and cluster analysis: B.R., S.-Y.C. and G.P.N.; RNA isolation and library preparation: B.R., K.J.P., R.S. and A.V.; sequencing: J.O., G.M. and N.F.N.; sequencing analysis and development of analytical tools: M.K., J.S., A.M.N. and S.R.Q.; immunological assays: B.R.; microscopy and experimental design: B.R., D.M.C., K.J.I., D.N.C. and A.V.; electron microscopy: L.M.; BHF localization to the membrane and serum validation: J.M.T., B.R. and A.V.; transplantation: B.R.; writing of manuscript: B.R., A.V., M.K., T.R., K.J.P. and I.L.W.; technical support and conceptual advice: N.F.N., D.M.C., A.M.N., T.R., G.P.N., S.R.Q., A.V. and I.L.W.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0783-x>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0783-x>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to B.R., I.L.W. or A.V.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Colony development, labelling and allorecognition assays. The life cycle of *B. schlosseri* includes both sexual and asexual reproduction pathways. Sexual reproduction starts with fertilization and progresses through classic embryonic stages into a tadpole larva with chordate characteristics such as a tail, notochord, neural tube and striated musculature. Upon hatching, the motile tadpole settles on a substrate and metamorphoses into an oozoid, with a sessile body plan (Fig. 1a). The oozoid begins a cyclical budding process of asexual reproduction, forming a colony of genetically identical zooids and buds (Fig. 1b). The colonial individuals are united under a single gelatinous tunic by a network of blood vessels, which terminate in sausage-shaped protrusions (ampullae; Fig. 1a–c). Throughout adult life, *B. schlosseri* regenerates its entire body every two weeks. This stem-cell-mediated cycle of development includes the formation of all body organs including the heart, respiratory system, digestive system and neural complex. Sexual reproduction commences when the ovary and testis are formed.

Mariculture procedures have previously been described³⁵. In brief, *B. schlosseri* colonies were collected from the marina in Monterey, California. Individual colonies were tied to 3 × 5-cm glass slides and placed 5 cm away from and opposite to another glass slide in a slide rack. The slide rack was placed into an aquarium, and within a few days the sexually reproduced tadpoles hatched, swam to the settlement slide, and metamorphosed into the adult body plan. Single colonies were then transferred to individual slides and used in the experiments.

When two genetically distinct colonies or individuals encounter one another, they can either fuse to form a chimaera with a common vasculature, or reject one another^{2,3,7} (Fig. 1c, Supplementary Video 2). This self–nonself recognition process is controlled by *BHF*⁷. Fusion requires at least one shared *BHF* allele: upon fusion, circulating somatic and germline stem cells from both colonies compete for dominance of the somatic and germline organs and also contribute to the formation of new buds^{4,5}. Additionally, the buds in one of the chimeric partners often fail to develop⁸. This developmental failure is an immune-cell-based rejection of bud cells that operate within a *BHF*-histocompatible chimaera. The process involves inflammation and recruitment of cytotoxic cells, and is thus comparable to chronic transplant rejection in mammals⁷.

Juvenile colonies were labelled with distinct fluorescent stains (CellTrace CFSE Green, Invitrogen #C34554; CellTrace FarRed, Invitrogen #C34572) and positioned in controlled fusion/rejection reactions (Fig. 3d). For allorecognition assays, colonies were labelled using CFSE dye (5 mM stock solution) and Far Red dye (1 mM stock solution) in a dilution of 1 µl of dye to 1 ml of filtered seawater. Naive colonies were then bathed in this solution for 60 min, allowed to recover then washed and placed in µ-dishes (ibidi µ-Dish 50 mm-low uncoated) for observation. Images were obtained using a confocal fluorescence microscope (LSM700 Axio Observer.Z1, Zeiss).

FACS analysis. For cell isolation, colonies were starved for at least 24 h before dissociation with a fine blade into cell suspensions. Cells were filtered through a 40-µm mesh using a sterile 1-ml syringe pump (a similar procedure to murine spleen dissociation), washed and collected in staining medium: 3.3 × PBS, 2% FCS and 10 mM Hepes. No enzymatic dissociation was used and the presence of FCS was enough to prevent cell aggregation. After gating on propidium iodide (PI), negative cells (using 2D plots owing to natural fluorescence of *B. schlosseri* cells) were analysed using a size–forward scatter (FSC) and granularity–side scatter (SSC) panel on a log scale using BD FACS Aria-II. We isolated 11 cell populations based on their size (FSC), granularity (SSC) and natural auto-fluorescence (Extended Data Fig. 1b). These purified populations were analysed according to morphology by light microscopy and defined according to previous cellular work characterizing *B. schlosseri* cell populations^{3,26}. We then screened a large assortment of antibodies and reagents using mass cytometry (CyTOF) and FACS, to identify cell surface markers that showed differential binding to distinct cell populations (Extended Data Tables 1, 2a, Extended Data Fig. 2). The cluster analysis programs SPADE and viSNE were applied to create a differentiation panel of the resultant screened markers (Extended Data Fig. 1a). Using anti-CD49d, anti-CD57, anti-BHF (a mouse anti-serum that was developed against the BHF protein), concanavalin-A, and alkaline phosphatase, together with size, granularity, and natural auto-fluorescence, we defined a total of 34 cell populations—24 of these are the hierarchical endpoint populations of our FACS gating strategy (Extended Data Fig. 1c, d). CP17 was composed of dead cells and cell debris, and was therefore excluded.

Thirty-four gated populations were analysed after cell labelling. Five million cells were suspended in 200 µl of staining medium before FACS analysis: alkaline phosphatase (AP) Live Stain (Life Technologies A14353) 1 µl, CD49d PE-Cy7 (BioLegend 304313; Clone 9F10) 1 µg, CD57 Pacific Blue (BioLegend 322316; clone HCD57) 0.25 µg, concanavalin-A (ConA) AlexaFluor-633 (Sigma) 2 µg, mouse anti-BHF serum 1:100 and anti-mouse secondary Cy5-Cy7 (SantaCruz)

(Extended Data Tables 1, 2a). The specific excitation laser and optical filter for emission measurements were as follows (excitation laser in nm and filters stated as long pass (LP) and band pass (BP)): 488 nm (505LP 530/30BP)–AP, 488 nm (755LP 780/60BP)–CD49d, 405 nm (450/50BP)–CD57, 633 nm (660/20BP)–ConA, 633 nm (755LP 780/60BP)–BHF. Cells were sorted into 18-well flat µ-slides coated with poly-L-lysine (ibidi) for live imaging.

Anti-BHF serum. We created the mouse anti-BHF serum with Thermo Scientific Pierce Protein Biology. The nucleotide sequence that was used was for the *BHF* (759 bp) α-allele, a *BHF* allele that is common in our mariculture facility:

```
ATGGTGCACGATACCGAGCAATTGCTCGCCCAAGGTCACCACGAAG
AGGAAACCGAATGTGGGAAATACGGAAAGCTGCCGGAGAAAGGCAGC
GAATGCAAGAAACATGGCATATTTTGCC GAATCCTGACTGCGTTA
CATTTGAAGAAGAG GAGAACTGAACACGATCATCAAAAGTT
ACTGTCCGAGTCGCAAGAGCATCTCGACGCTTCGACAAAGAAGACCAA
GAAAAAAGCCAAGAAGGACAAACGCAAAACAAACCGCCCAAGAAAG
ATTCAGAACTAGCAAGCCCGCTCAGACCACGATTTCAAGACTCCCAT
CAAA CAGAAACAATAACAACGCAACAGCTTCGCTACCACCTACGAA
AAATTCGACAACGACTCACTGTGCTCTGTGCACTTAATACGGGTCGATA
TCGAATTTTGGGACATGGAGAACGAACAGTCGACCAACTACCCACG
AAATCCTGGAATCTGTTTCATATGTACGGCGACGATCGGTTTCGGCGAAC
GCCTCATTTAGTAGCCCCAAAATAATACGCCCGCTTGGACGAAAGC
AACGGTCCGAGTCCACGGCGCTGGGGAGTACTTAAAGCATCAGTGG
AAGGGTCAGGGGGCCAAAAAAGCGCGCAAGAGAATTGCAACTGTGAT
GAAGGTACTTGGCAATCTCTACAAGCGGGCGCTAGATCGCAACAGC
TTTTTTGAACCTCAGGGTGCAGGTGTCGGCGGCCCTGTACAAAACAG
GAGATAA
```

The *BHF* sequence was cloned into a pET23b vector including a C-terminal His tag, and produced in DH5a cells. Validation of binding of mouse serum to BHF was done by Thermo Scientific Pierce Protein Biology using ELISA. We used FACS to validate that the anti-BHF mouse serum bound to *Botryllus* live cells. Labelling of live cells (propidium iodide negative) by the anti-BHF serum shows that BHF is present on the membrane (Extended Data Fig. 2e). The localization of the BHF to the membrane has previously been described³⁶. Our assumption is that if a serum can recognize BHF on the membrane it will be able to block its interaction with a recognition receptor.

CyTOF mass cytometry screening. CyTOF and its cluster analysis programs SPADE and viSNE were used to screen a large assortment of antibodies against human cell-surface antigens that would cross-react with *B. schlosseri* cells and could potentially differentiate cellular populations (Extended Data Table 1, Extended Data Fig. 2). The cluster analysis programs SPADE and viSNE, which were developed to analyse CyTOF data³⁷, were used to create a differentiation panel of the resultant screened markers (Extended Data Fig. 1).

For the screen of 49 antibodies, live cells were labelled for 30 min on ice in staining medium (3.3 × PBS, 2% FCS and 10 mM Hepes) with different antibodies labelled with elemental isotopes (Extended Data Table 1), followed by two washes with staining medium. After washing, cells were fixed in 4% PFA in 1 × PBS, washed once with staining medium and then incubated at room temperature for 20 min in an iridium-containing DNA intercalator (Fluidigm) in 1.5% paraformaldehyde in 1 × PBS. Prior to measurement on a mass cytometer, cell samples were washed once with staining medium and twice with water³⁷ (Fluidigm). Analysis of the FCS files were done with Cytobank and FlowJo. Antibodies with positive signals were validated by FACS as summarized in Extended Data Table 1.

Tissue dissection, RNA extraction, purification and transcriptome sequencing.

FACS-sorted cell population. We used a published protocol³⁸ to extract RNA from FACS-sorted cells ($n = 24$). Twenty-three endpoint populations were sequenced, in addition to CP2, which was composed of all live cells. CP17 was composed of dead cells and cell debris, and therefore was excluded from the population analysis. Twenty thousand cells of each sequenced population were sorted in 750 µl Trizol-LS (Invitrogen #10296010) using a 100-µm nozzle. Cells were vortexed and incubated for 10 min at room temperature before freezing at -80°C . After thawing, cells were washed with chloroform, and RNA was isolated according to the manufacturer's directions, with minor modifications. A linear polyacrylamide (LPA) carrier (Sigma Aldrich # 56575-1ML) was added to enhance recovery of RNA, followed by DNase I treatment per the manufacturer's instructions (Qiagen RNeasy micro kit, # 74004).

Whole-colony samples. We used a published protocol⁶ to extract RNA from whole colonies ($n = 34$). The 34 whole-colony samples included all tissues in the colonies. **Endostyle samples.** Insulin syringes were used to dissect endostyles ($n = 10$). The endostyle is a long glandular groove extending medially at the ventral face of the zooid's branchial sac along its anterior–posterior axis²³. The endostyle is immersed in blood cells flowing through the large subendostylar sinus and other sinuses and is innervated by the main and lateral subendostylar nerves^{39,40}. The endostyle epithelium consists of eight distinct anatomical zones, each defined by a specific gene expression profile^{6,16,21,22}. The ten endostyle samples taken for RNA-seq

included the endostyle epithelial cells, cells circulating in the subendostylar sinuses and the main and lateral subendostylar nerves, which extend along the endostyle anterior–posterior axis. Dissected endostyle samples were flash-frozen in liquid nitrogen to minimize RNA degradation and stored at -80°C . Using a mechanized Konte tissue grinder and pestle, samples were homogenized in the presence of lysis buffer (Qiagen RNeasy Microkit #74004), and total RNA was extracted following the manufacturer's protocol. The resultant RNA was cleaned and concentrated (Zymo Research RNA Clean and Concentrator-5, R1015) and analysed using an Agilent 2100 Bioanalyzer for quality analysis before library preparation. cDNA libraries were then prepared from high-quality samples (RNA integrity number (RIN) > 8) using Ovation RNA-seq v2 (Nugen). Size selection was performed before barcoding using Zymo Research Select-a-Size DNA Clean and Concentrator Kit (D4080). Libraries were barcoded using NEBnext Ultra DNA Library Prep Kit Master for Illumina (New England Biolabs, E7370S) and NEBNext Multiplex Oligos for Illumina (New England Biolabs, E6609S). Barcoded library samples were then sequenced on an Illumina NextSeq 500 (2×150 bp, producing an average of 15 million reads per cell population).

Cytotoxicity and phagocytosis assays. Three different ex vivo phagocytosis assays were used to identify the *B. schlosseri* myeloid lineage phagocytic cell populations: (i) phagocytosis of fluorescent beads; (ii) phagocytosis of a fluorescently labelled marine bacterium (*Vibrio diazotrophicus*); and (iii) allogeneic phagocytosis, during which we tested the capability of cells from different colonies to engulf allogeneic cells (Extended Data Fig. 7). FACS was used to identify the phagocytic cells in each of the ex vivo assays and to track their cell populations. Confocal microscopy and ImageStream analysis were used to confirm engulfment (Extended Data Fig. 7).

Bacteria and beads phagocytosis assays. Cells (10^5 cells/ $200\ \mu\text{l}$) were incubated overnight at 18°C in a 2:1 ratio of beads:cells using Fluoresbrite YG Carboxylate Microspheres $1.00\ \mu\text{m}$ (Polysciences). To measure phagocytosis of bacteria, the marine bacterium *V. diazotrophicus* was heat-inactivated at 95°C for 5 min, and labelled with Alexa Fluor 647 (Invitrogen #A20006). The analysis of cells positive for beads or bacteria was done by flow cytometry using two fluorescent channels—the green channel was used to detect beads and the far-red channel was used to detect bacteria (Extended Data Fig. 7). FlowJo V10 (FlowJo) was used to analyse the flow cytometry data.

Phagocytosis of allogeneic cells. The labelling and incubation of cells were done as described for the cytotoxicity assay below. Phagocytosis was defined as double-positive cells in the FACS plots of the two labelling markers. Owing to the natural fluorescence of *B. schlosseri* cells, the level of double-positive cells in wells with separation of each one of the labels was reduced from the double-positive in the experimental wells (about 5% background). To validate phagocytosis, cells were sorted and observed by confocal microscopy. To compare cytotoxicity to phagocytosis, the optimal ratio (1:1) for allogeneic phagocytosis was prepared and compared to cytotoxicity at the same ratio, using 10^5 cells of each labelled group per well.

We used a FACS-based cytotoxicity assay^{8,41} to measure killing of cells in vitro. Isolated *B. schlosseri* cells were labelled using CFSE ($5\ \mu\text{M}$; Life Technologies) and Far Red dye ($1\ \mu\text{M}$; Life Technologies) for 30 min at 18°C to distinguish effector and target cells and washed twice in staining medium: $3.3\times$ PBS, 2% FCS and 10 mM Hepes. Cells were incubated overnight in 96-well U-shaped plates at 18°C at different effector:target ratios in staining medium. After adding propidium iodide to wells to test cell viability, cells were analysed by FACS. The target or the effector cells were gated on 2D analysis using the labelling dyes (CFSE and Far Red). Spontaneous lysis was measured as the percentage of propidium iodide-positive cells in gated target cells in wells without effector cells, and sample lysis was quantified as the percentage of propidium iodide-positive cells in gated target cells. Specific lysis was calculated as follows: specific lysis % = $100 \times ((\text{sample lysis} - \text{spontaneous lysis}) / (100 - \text{spontaneous lysis}))$. The gating of FACS analysis was on 2D plots owing to the natural fluorescence of *B. schlosseri* cells. For BHF blocking assays, anti-BHF polyclonal mouse serum was used at 1:200, or mock serum as control. Colonies with known fusion or rejection outcomes from our mariculture facility were used to measure allogeneic cytotoxicity and phagocytosis. Cytotoxicity and phagocytosis assays were done in triplicate.

Cell transplantation. *B. schlosseri* colonies have the ability to fuse spontaneously, thereby facilitating natural 'transplantation' or parabiosis assays between genetically compatible but distinct colonies (those sharing a BHF allele) without requiring radiation-induced elimination of the host's stem cells^{5,6,16}. Furthermore, based on this feature, cells can be sorted and transplanted between compatible *B. schlosseri* colonies. The differentiation potential of transplanted cells can be tested using a pigmentation-based assay, using distinctly pigmented donor and recipient colonies (for example, blue versus orange)^{5,6}. The colours of the pigment cells in *B. schlosseri* are genetically determined^{42,43}. Pigment cells, identified as CP8, circulate in the vasculature and are linked to the haematopoietic lineages by gene expression (Extended Data Fig. 4b). Additionally, the transparent body allows real-time monitoring and tracing of labelled cells by time-lapse microscopy⁶. Using

these methods and a FACS-based differentiation assay developed for this study, we assayed the differentiation potential and homing sites of *B. schlosseri* candidate HSC (cHSC) populations. The experiment of rescuing lethally irradiated recipients was not performed because those experiments were not shown to work in the *B. schlosseri* model⁴⁴.

Cells were sorted directly into staining medium ($3.3\times$ PBS-based; 75% of final volume) to minimize cellular stress using a $100\text{-}\mu\text{m}$ nozzle. Following sorting, cells were labelled with DiD or DiO membrane dye (Life Technologies) to visualize or identify the transplanted population, and suspended in a final concentration of 10^5 cells per μl . We injected 10^5 cells into recipient colonies near the zoid's heart, using a microinjector (Narishige) as described⁶.

Two transplantation assays were used to measure the ability of a specific cell population to differentiate into other cell types: (i) a pigment cell-based differentiation assay (donor chimerism) (Fig. 2b, c), and (ii) a FACS analysis-based differentiation assay (Fig. 2d–f).

The cHSC populations transplanted include CP25, CP33 and CP34. As a control we used CP18 (includes CP19, CP20 and CP21), which shares a molecular signature with the pigment cell population and is located close to the pigment cell populations on the FACS plots (Extended Data Fig. 1). The analysis of the pigment cell-based differentiation experiment (Fig. 2b, c) was done in a single-blinded manner 20 days after transplantation. $n = 42$ transplanted colonies; 16 cHSC, 15 control population and 11 uninjected.

In the FACS analysis-based differentiation assay (Fig. 2d–f), the cHSC populations (CP25, CP33 and CP34) and the CP18 control population (CP19, CP20, CP21) were labelled with DiO and transplanted into compatible colonies. Three weeks after transplantation, transplanted colonies were pooled for analysis by FACS into two pools from each experimental group, with at least five transplanted colonies in each pool. The percentage of cells remaining in each of the original gates on the FSC and SSC was measured, CP9 for CP18 (CP19, 20, 21) and CP10+7 for the cHSC population (CP25, 33, 34) based on the FACS gating hierarchies in Extended Data Fig. 1d. In addition, we estimated cell proliferation by measuring levels of DiO fluorescence 3 weeks after transplantation of labelled cells (Extended Data Fig. 5a).

For localization assays, two sets of experiments were performed: a total of 14 colonies labelled with CFSE on ibidi 50-mm uncoated μ -dishes were transplanted with DiD-labelled cells (Fig. 2g, Extended Data Fig. 5b). cHSC DiD-labelled cells were transplanted into six colonies, four colonies were transplanted with DiD-labelled CP3, and four colonies were not injected and used as a control for natural fluorescent background. Focusing on known *B. schlosseri* stem-cell niches (the subendostylar sinus and the cell islands^{6,16}), we tested the ability of transplanted cHSCs and control cells (CP3) to migrate to these sites. An additional channel was used to validate autofluorescent cells. For the subendostylar sinus (the endo-niche) we detected no transplanted cells for CP3 (0/4) or the uninjected colonies (0/4), whereas 5/6 colonies were positive for cHSC cells, showing significant localization of the cHSCs to the subendostylar sinus; $P = 0.048$, Fisher's exact test, two-tailed. However, in the cell islands, although 4/4 were positive with CP3, 5/6 positive to cHSC, and 2/4 were positive in the uninjected colonies, there is a high concentration of autofluorescent cells in the cell islands.

Histology and immunohistochemistry. For haematoxylin and eosin (H&E) staining, cells were incubated overnight on glass slides coated with poly-L-lysine (Sigma) and fixed with 4% PFA in 0.1 M MOPS in 0.5 M NaCl, pH 7.5 for 10 min and washed with PBS. The slides were stained with Harris haematoxylin for 5 min and 2% eosin Y for 1 min (Extended Data Fig. 1e).

Immunohistochemistry was based on a published method⁴⁵. The labelling was done using rabbit anti phospho-histone H3 (pHH3; Cell Signaling, 9701) at a concentration of 1:500 (Extended Data Fig. 5d).

Electron microscopy. Colonies were fixed overnight in 1.5% glutaraldehyde buffered with 0.2 M sodium cacodylate, 1.6% NaCl, pH 7.4. After washing in buffer and post-fixation in 1% OsO₄ in 0.2 M cacodylate buffer plus 1.6% NaCl, specimens were dehydrated and embedded in Araldite. Sections were counterstained with toluidine blue, and for electron microscopy sections were stained for contrast with uranyl acetate and lead citrate. Micrographs were taken with a FEI Tecnai G2 electron microscope (operating at 100 kV) (Extended Data Fig. 5c).

Differential expression analysis. Determination of gene counts was performed using a Snakemake⁴⁶ pipeline. An outline of the steps is as follows: (i) low-quality bases and adaptor sequences were removed using Trimmomatic⁴⁷ (version 0.32); (ii) overlapping paired end reads were merged using FLASH⁴⁸ (version 1.2.11); (iii) reads were aligned to the UniVec Core database using Bowtie2⁴⁹ (version 2.2.4) to remove biological vector and control sequences; (iv) reads were aligned to the *B. schlosseri* transcriptome with BWA⁵⁰ (mem algorithm, version 0.7.12); and (v) aligned reads were sorted and indexed using SAMtools, PCR duplicates removed using PICARD (MarkDuplicates tool, version 1.128) and then transcript level counts directly counted from the BAM file.

Differential expression was analysed using edgeR⁵¹. In detail, the gene counts were compiled into a tabular format and loaded into R. Genes were retained with at

least five counts per million in at least 80% of the smaller number of the compared samples. A simple model was used to compare the two sets of populations, with *P* values adjusted using the Benjamini–Hochberg method to produce a false discovery rate (FDR). FDRs below 0.05 were called as being differentially expressed. The comparisons between cell populations were performed in a one-versus-all approach, followed by selective aggregating of similar populations. Initially each population was compared to all others (except for CP2, which was an aggregate of all cells). Then all sets of two populations were compared to the remaining populations, with the best two being grouped together. The metric used was: if *A* is the set of genes found to be differentially expressed in population *A* versus all others, and *n*(*A*) is the number of genes in that set, then find the maximum of *n*(*AB*) – *n*(*A* or *B*). This attempts to find those populations that when grouped together are more distinct from all others compared to the populations individually. After this grouping was performed multiple times, the maximum was no longer above zero, and a new metric was used. This method was used to find the two groups of populations that maximized $n(AB)^2/(n(A \text{ or } B) + 1)$. This aimed to find the populations that fractionally were the most similar. The mergings are as follows: CP33 + CP34, CP3 + CP23, CP8 + CP20, CP21 + CP22, CP24 + CP26, CP25 + CP33 + CP34, CP12 + CP14, CP5 + CP32, CP29 + CP36, CP19 + CP31, CP15 + CP16, CP30 + CP35. The number of characteristic, highly expressed genes found per population(s) relative to the other populations varied from 0 to 2,229 (mean 232), with only 23% possessing homology to mouse or human genes and the remaining 77% of these genes being mostly *B. schlosseri*- or tunicate-specific.

For visualization, the 250 genes with the largest weights in the first 11 principal components (explaining 90% of the variance in mean-adjusted log-transformed gene counts) were used to cluster the different cell populations in a heat map (Extended Data Fig. 3a). Many of the populations have visually similar gene expression patterns (for example: CP3, 5 and 23; and CP25, 33 and 34), and in general, samples that were adjacent in FACS space are also similar in expression space (Extended Data Fig. 3b, 67% of neighbours in agreement), indicating a correlation between gene expression profile and morphology and marker expression.

Gene domain finding. The protein sequences of CP31-associated non-homologous genes (*n* = 52) were saved to individual FASTA files. These were processed using the python-based RESTful client for InterPro5 (version 5.28–67.0) on ENSEMBLE's website using the default options to search for annotated domains within ascidian-specific genes that dominate the *Botryllus schlosseri* morula cells (Supplementary Table 5). Unannotated gene domains revealed: transmembrane domains (*n* = 7), cell interaction and sugar binding domains (*n* = 4) and signal peptides (*n* = 3). This analysis also revealed genes with toxin and peptidases domains (g6777, g43113), hydrolase (g61144) and complement and coagulation domains (g6900, g69753). Moreover, g15971 contained a C-type lectin domain, a domain contained in human natural killer cell receptors (Supplementary Table 5). **Geneset Activity Analysis on Gene Expression Commons.** Gene Expression Commons (<https://gexic.riken.jp>) provides gene expression dynamic ranges and thresholds to distinguish active expression from inactive expression for each gene by computing a massive amount of publicly available microarray data¹⁵. Once a user submits raw microarray data of a particular cell type, the Gene Expression Commons returns gene expression activity referred to the dynamic range for each gene, and if the expression level exceeds the threshold, the gene is labelled an active gene. In the Weissman laboratory, we purified and generated microarray data for each step of adult mouse haematopoiesis as well as several stromal cell types. A gene expression profile of each cell type becomes public domain (or available to the public) through the Gene Expression Commons platform.

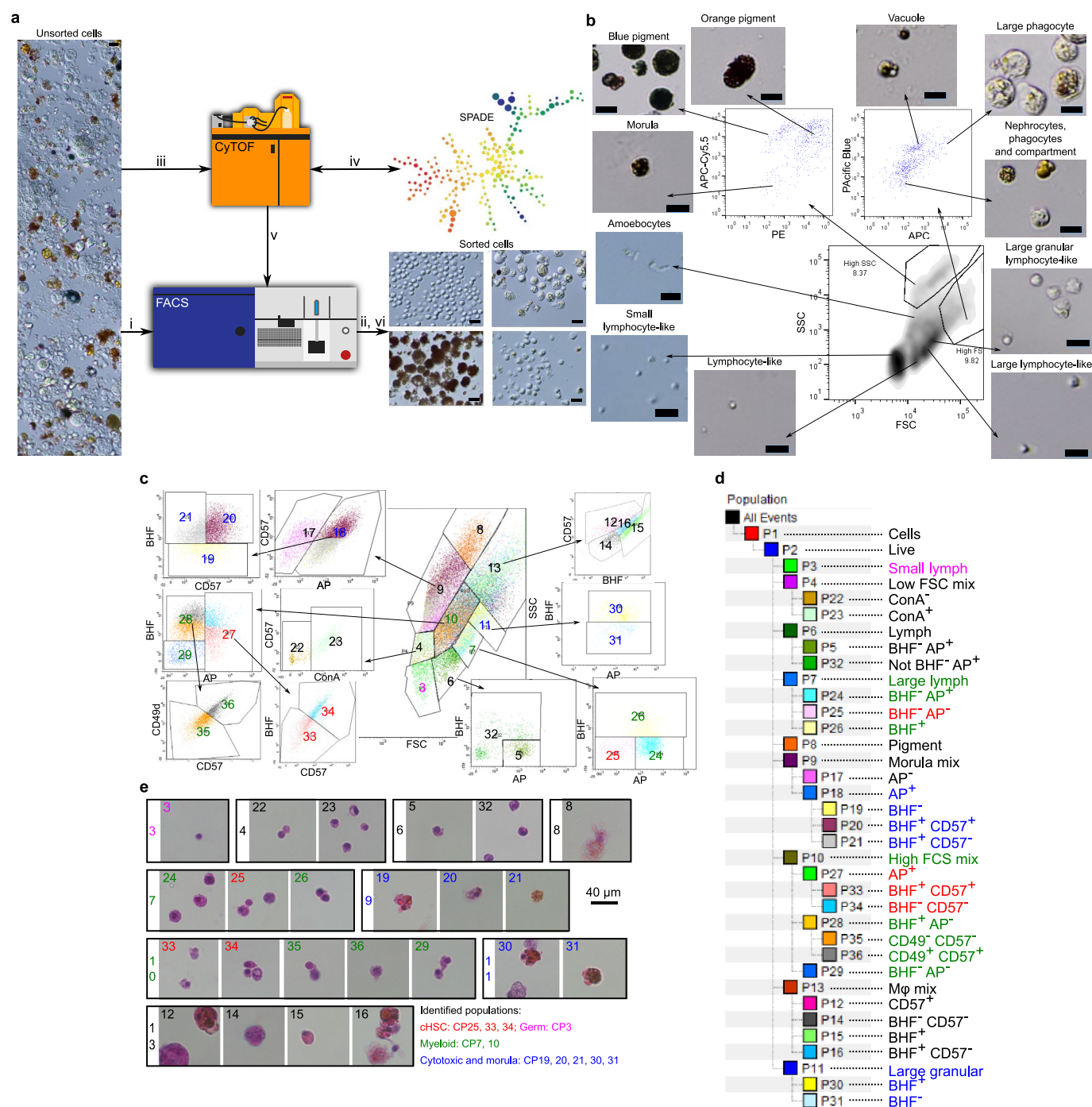
For this project, we developed a new function on the Gene Expression Commons platform entitled Geneset Activity Analysis. A user can create a Geneset, a set of genes of interest. Then the significance of the association between a given gene set and a set of active genes in the cell type is examined by Fisher's exact test. **Reporting summary.** Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Sequencing data can be found on the NCBI Sequence Read Archive under accession PRJNA414486. RPKM values of gene expression and differential expression

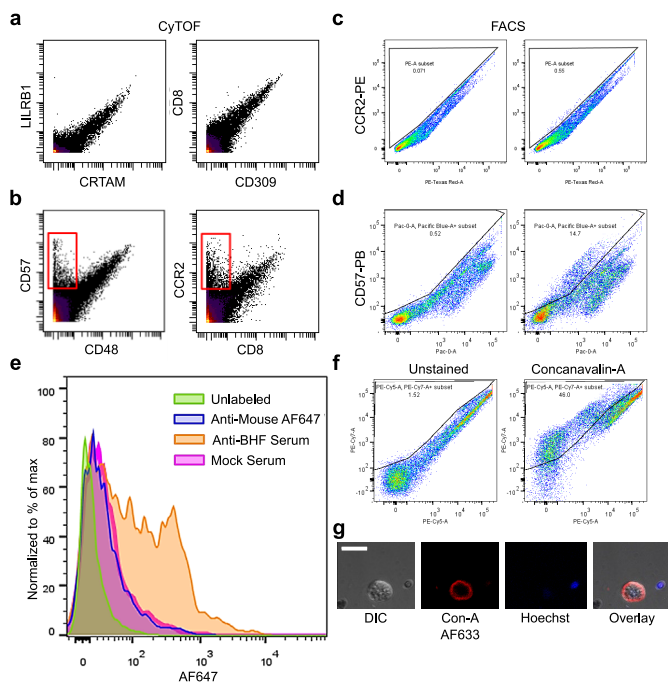
analysis results are shown in Supplementary Table 1. All other relevant data are available in the manuscript.

35. Boyd, H. C. et al. Growth and sexual maturation of laboratory-cultured Monterey *Botryllus schlosseri*. *Biol. Bull.* **170**, 91–109 (1986).
36. Taketa, D. A. & De Tomaso, A. W. *Botryllus schlosseri* allorecognition: tackling the enigma. *Dev. Comp. Immunol.* **48**, 254–265 (2015).
37. Bendall, S. C. et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011).
38. Koh, P. W. et al. An atlas of transcriptional, chromatin accessibility, and surface marker changes in human mesoderm development. *Sci. Data* **3**, 160109 (2016).
39. Burighel, P. & Brunetti, R. The circulatory system in the blastozooid of the colonial ascidian *Botryllus schlosseri* (Pallas). *Italian J. Zool.* **38**, 273–289 (1971).
40. Zaniolo, G., Lane, N. J., Burighel, P. & Manni, L. Development of the motor nervous system in ascidians. *J. Comp. Neurol.* **443**, 124–135 (2002).
41. Edri-Brami, M. et al. Glycans in sera of amyotrophic lateral sclerosis patients and their role in killing neuronal cells. *PLoS ONE* **7**, e35772 (2012).
42. Sabbadin, A. & Graziani, G. Microgeographical and ecological distribution of colour morphs of *Botryllus schlosseri* (Ascidacea). *Nature* **213**, 815–816 (1967).
43. Cima, F. et al. Life history and ecological genetics of the colonial ascidian *Botryllus schlosseri*. *Zool. Anz.* **257**, 54–70 (2015).
44. Laird, D. J. & Weissman, I. L. Continuous development precludes radioprotection in a colonial ascidian. *Dev. Comp. Immunol.* **28**, 201–209 (2004).
45. Braden, B. P. et al. Vascular regeneration in a basal chordate is due to the presence of immobile, bi-functional cells. *PLoS ONE* **9**, e95460 (2014).
46. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
47. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
48. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
49. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
50. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
51. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
52. Wolff, L. & Humeniuk, R. Concise review: erythroid versus myeloid lineage commitment: regulating the master regulators. *Stem Cells* **31**, 1237–1244 (2013).
53. Rosental, B. et al. The effect of chemotherapy/radiotherapy on cancerous pattern recognition by NK cells. *Curr. Med. Chem.* **19**, 1780–1791 (2012).
54. Potts, K. S. & Bowman, T. V. Modeling myeloid malignancies using zebrafish. *Front. Oncol.* **7**, 297 (2017).
55. Moss, L. D. et al. Identification of phagocytic cells, NK-like cytotoxic cell activity and the production of cellular exudates in the coelomic cavity of adult zebrafish. *Dev. Comp. Immunol.* **33**, 1077–1087 (2009).
56. Wei, S. et al. The zebrafish activating immune receptor Nitr9 signals via Dap12. *Immunogenetics* **59**, 813–821 (2007).
57. Tang, Q. et al. Dissecting hematopoietic and renal cell heterogeneity in adult zebrafish at single-cell resolution using RNA sequencing. *J. Exp. Med.* **214**, 2875–2887 (2017).
58. Han, Q. et al. Characterization of lamprey IL-17 family members and their receptors. *J. Immunol.* **195**, 5440–5451 (2015).
59. Hirano, M., Das, S., Guo, P. & Cooper, M. D. The evolution of adaptive immunity in vertebrates. *Adv. Immunol.* **109**, 125–157 (2011).
60. Han, Y., Huang, G., Zhang, Q., Yuan, S. & Liu, J. The primitive immune system of amphioxus provides insights into the ancestral structure of the vertebrate immune system. *Dev. Comp. Immunol.* **34**, 791–796 (2010).
61. Rhodes, C. P., Ratcliffe, N. A. & Rowley, A. F. Presence of coelomocytes in the primitive chordate amphioxus (*Branchiostoma lanceolatum*). *Science* **217**, 263–265 (1982).
62. Muñoz-Chápuli, R., Carmona, R., Guadix, J. A., Macías, D. & Pérez-Pomares, J. M. The origin of the endothelial cells: an evo-devo approach for the invertebrate/vertebrate transition of the circulatory system. *Evol. Dev.* **7**, 351–358 (2005).
63. Lanot, R., Zachary, D., Holder, F. & Meister, M. Postembryonic hematopoiesis in *Drosophila*. *Dev. Biol.* **230**, 243–257 (2001).
64. Meister, M. & Lagaveux, M. *Drosophila* blood cells. *Cell. Microbiol.* **5**, 573–580 (2003).
65. Nakayama, K., Nomoto, A. M. & Nishijima, M. Morphological and functional characterization of hemocytes in the giant clam *Tridacna crocea*. *J. Invertebr. Pathol.* **69**, 105–111 (1997).

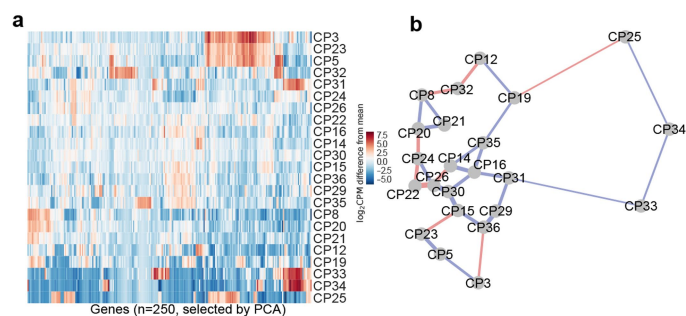


Extended Data Fig. 1 | *B. schlosseri* cell sorting workflow. **a**, Outline of cell purification process. Unsorted cells (light microscopy) are loaded into a FACS (i) and sorted, and this is followed by morphological observation (ii). Cells were labelled with diverse markers and screened using CyTOF (iii) for differential labelling. On the basis of SPADE cluster analysis (iv) markers were selected for FACS gating (v) before a final sort was performed (vi; c). **b**, Sorting based on FCS/SSC in the lower panel, and natural fluorescence in the upper panels. The analysis is after gating propidium iodide (PI)-negative cells (live cells). The specific excitation laser and optical filter for emission measurements were as follows (excitation laser in nm and filters stated as long pass (LP) and band pass (BP)): 488 nm for FSC and SSC, 488 nm (550LP 575/25BP)- PE, 405 nm (450/50BP)-Pacific Blue, 633 nm (660/20BP)-APC, 633 nm (690LP 710/50BP)-APC-Cy5.5. Nomenclature based on published work^{3,26}. Experiment was performed three times. Scale bars, 20 μ m. **c**, Sorting

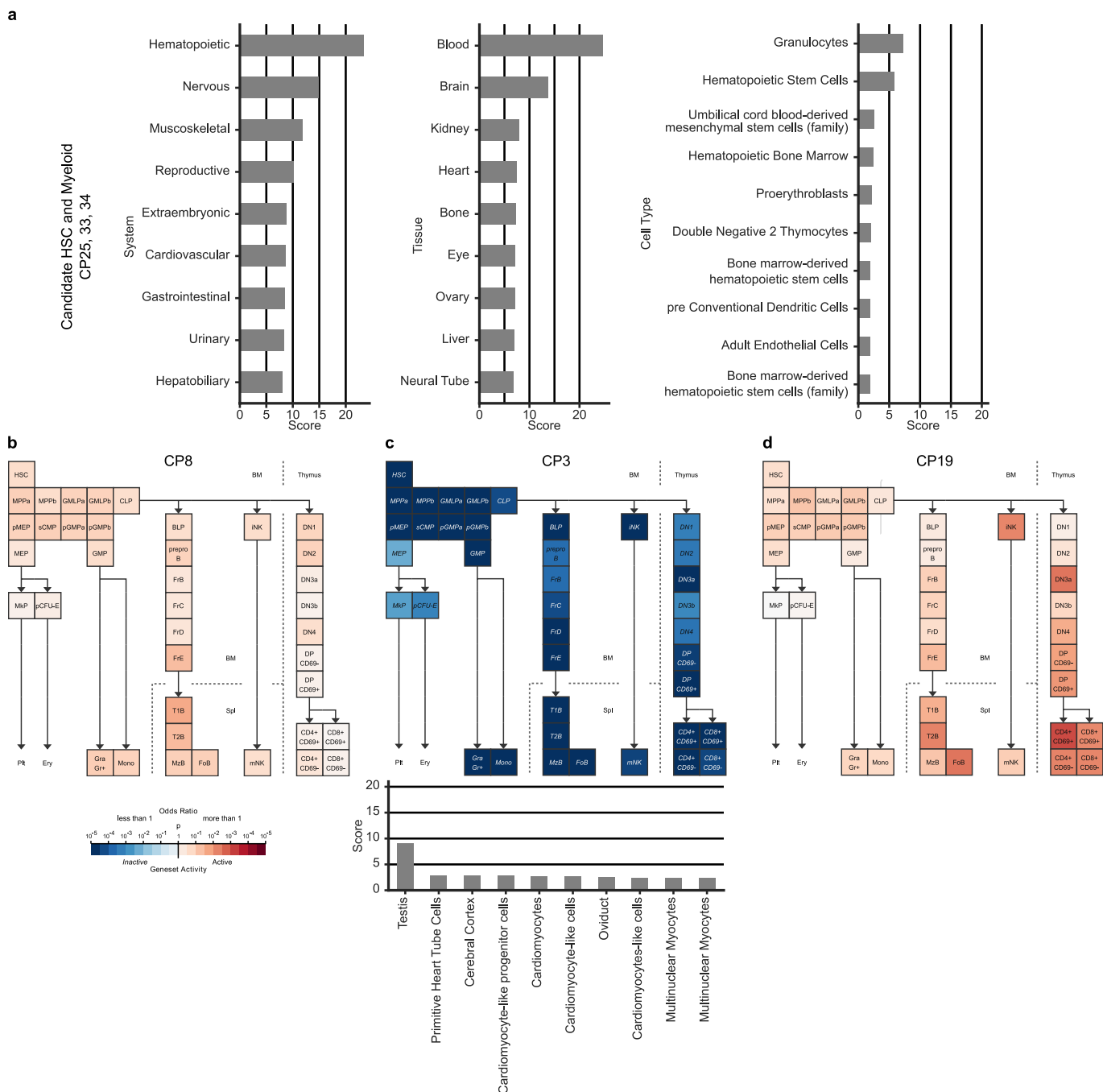
panels of 34 cell populations using FSC, SSC (central panel), CD49d, CD57, concanavalin-A (ConA), BHF and AP, after gating PI-negative cells (live cells). Central panel is FCS/SSC, from which additional populations were differentiated. The specific excitation laser and optical filter for emission measurements were as follows: 488 nm (505LP 530/30BP)-AP, 488 nm (755LP 780/60BP)-CD49d, 405 nm (450/50BP)-CD57, 633 nm (660/20BP)-ConA, 633 nm (755LP 780/60BP)-BHF and AP, after gating PI-negative cells (live cells). Central panel is FCS/SSC, from which additional populations were differentiated. The specific excitation laser and optical filter for emission measurements were as follows: 488 nm (505LP 530/30BP)-AP, 488 nm (755LP 780/60BP)-CD49d, 405 nm (450/50BP)-CD57, 633 nm (660/20BP)-ConA, 633 nm (755LP 780/60BP)-BHF and AP, after gating PI-negative cells (live cells).



Extended Data Fig. 2 | Screen for differentiation markers of cell populations for FACS-based sorting. **a, b,** Examples of antibodies screened by CyTOF mass cytometry with *B. schlosseri* cells analysed by 2D mass spectrometry. The CyTOF screen was performed once. **a,** Examples of antibodies considered as nonspecific binders because different antibodies showed the same binding patterns. **b,** Examples of antibodies considered as specific binders because a cell population was bound by one antibody but not the other (red rectangle). **c, d,** Examples of validation screen by flow cytometry of antibodies with specific binding by CyTOF, done in 2D fluorescence excited by the same laser owing to autofluorescence. Negative FACS validation was done twice and positive three times. **c,** Negative validation of CCR2 labelled with PE. **d,** Positive validation of CD57 labelled with Pacific Blue. **e,** Flow cytometry of live *B. schlosseri* cells labelled with anti-BHF mouse serum. Anti-mouse Alexa Fluor-647 was used as a secondary antibody. **f,** Example of positive differential labelling by the lectin concanavalin-A in PE-Cy7 by FACS. The experiment was performed three times. **g,** Confocal imagery of membrane concanavalin-A labelling with Alexa Fluor-633 in red and Hoechst DNA labelling in blue. The experiment was performed once. Scale bar, 20 μm .



Extended Data Fig. 3 | *B. schlosseri* cell population clustering based on transcriptome analysis. **a**, We used 250 genes with the largest weights in the first 11 principal components (explaining 90% of the variance in mean-adjusted log-transformed gene counts) to cluster the different cell populations in a heat map. **b**, Transcriptome sequencing of *B. schlosseri* cell populations compared to FACS analysis. 2D projection of the distances between transcriptomes of cell populations based on all differential genes. Lines are drawn between the nearest two neighbours. Blue, FACS adjacency of the populations in the differentiation panel; red, genetic level proximity not predicted by FACS panel adjacency. The proximities of twenty (of thirty) genetic level cellular populations were predicted by FACS. Widths of lines are inversely proportional to distances.

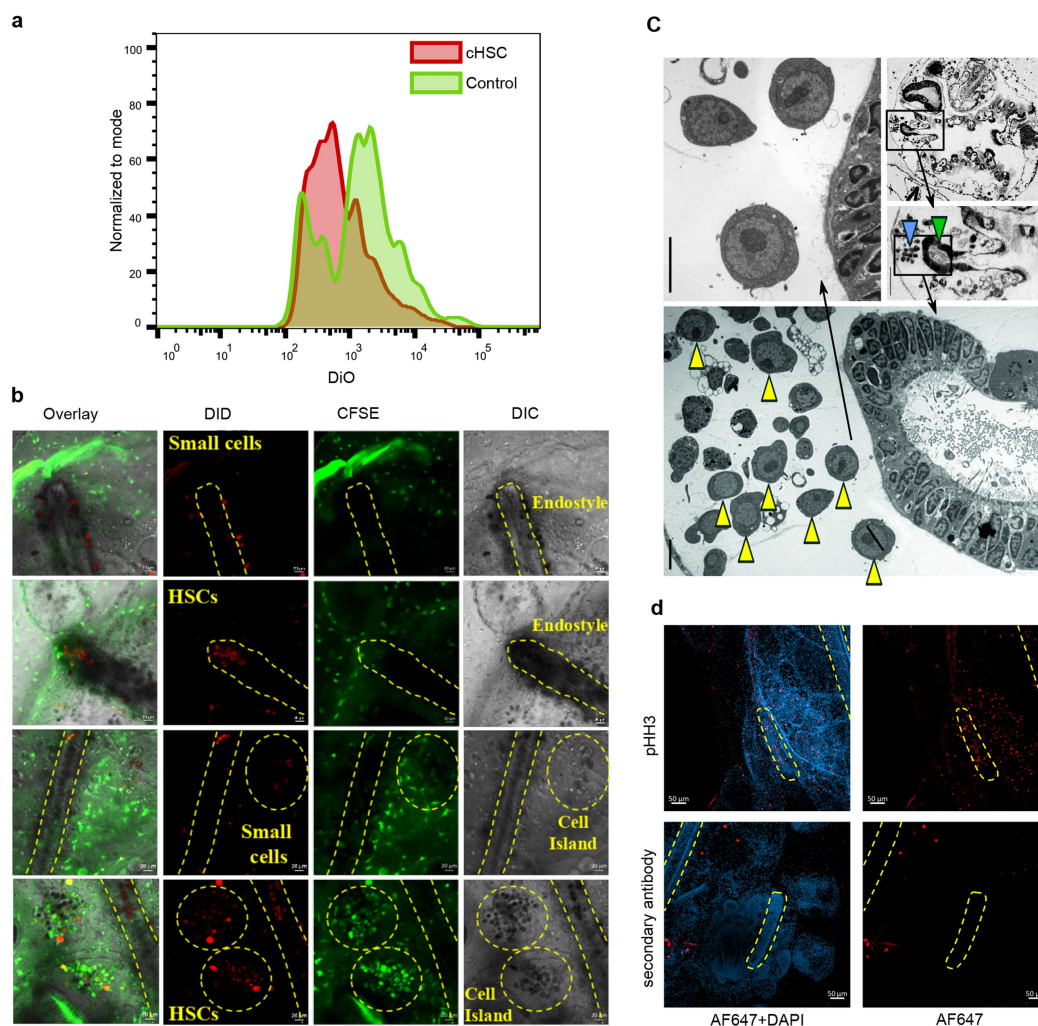


Extended Data Fig. 4 | Gene expression in *B. schlosseri* cell populations.

a, Enrichment scores of the top nine systems (left), nine tissues (centre) and ten cell types (right) of annotated genes upregulated in CP33, CP34 and CP25 using the GeneAnalytics tool (compared to human). In systems the haematopoietic system has the highest score, in tissues the blood has the highest score, and within the cells, the granulocytes and HSC have the highest scores. **b–d**, Geneset Activity Analyses using the Gene Expression Commons tool on a mouse haematopoiesis model of different *B. schlosseri* cell populations. **b**, Analysis of CP8 (pigment cells) based on 12 significantly upregulated genes, showing that CP8 is part of the

haematopoietic system with gene activity resembling known cell types.

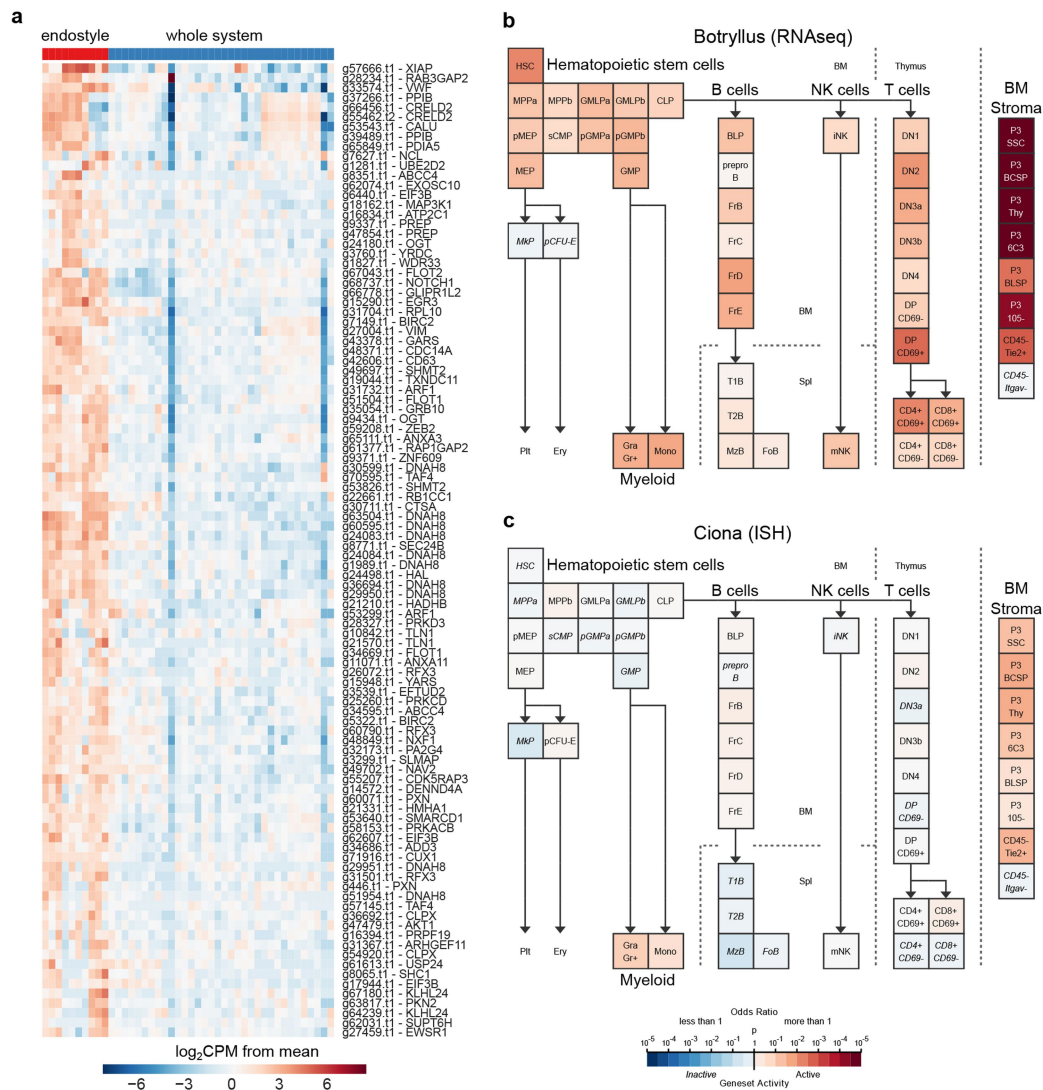
c, Analysis of CP3 (small cells) based on 235 significantly upregulated genes, showing that this population is not part of the haematopoietic system. On the bottom: enrichment scores of the top ten tissues using CP3 genes by GeneAnalytics tool compared to human; the highest score is for the testes, suggesting that this population is a gonadal population. **d**, Analysis of CP19 (enriched with morula cells) based on 96 upregulated genes ($P < 0.25$), showing that CP19 has gene activity resembling cells in the lymphoid lineage using Geneset analysis.



Extended Data Fig. 5 | Subendostylar sinuses as an HSC niche.

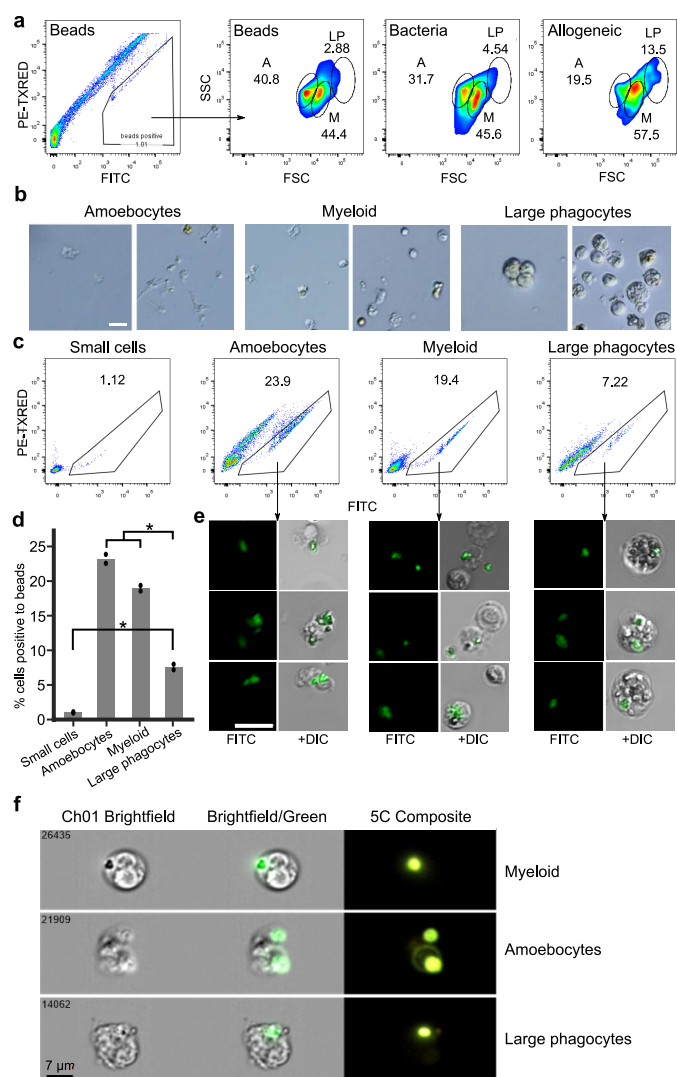
a, Reduction in DiO fluorescence suggesting cell proliferation, three weeks after transplantation. The experiment was performed once with two pools for each population from five animals each. **b**, Candidate HSC (cHSC) population and a control population (CP3) were isolated, labelled with DiD and transplanted into CFSE-labelled compatible colonies; in vivo tracing of transplanted cell migration was used to identify niches. There were no cells detected for CP3 (0/4) or the uninjected colonies (0/4) in the subendostylar sinus, whereas 5/6 colonies injected with cHSC cells showed significant localization of the cHSCs to the subendostylar sinus. $P = 0.048$, Fisher's exact test, two-tailed. Although in the cell islands 4/4 were positive with CP3, 5/6 positive with cHSC, and 2/4 positive in the

uninjected colonies, there are high levels of autofluorescent cells in the cell islands. Full image panels of Fig. 2g. **c**, Transverse sections of an adult zoid counterstained with toluidine blue (top two left) where the endostyle (green arrowhead) and endo-niche (blue arrowhead) are enlarged (scale bar, 30 μm). Electron microscopy section of the same animal's endostyle and endo-niche (right and bottom, enlarged). Yellow arrowheads indicate cells with haemoblast (HSC) morphology that are enriched within the endo-niche (scale bar, 5 μm). The experiment was performed three times. Full image panels of Fig. 2b. **d**, Immunohistochemistry with antibodies against phospho-histone H3 (PHH3), suggesting that there are mitotic cells in the endostyle region in the developing primary bud and also in the adult zoid endostyle.

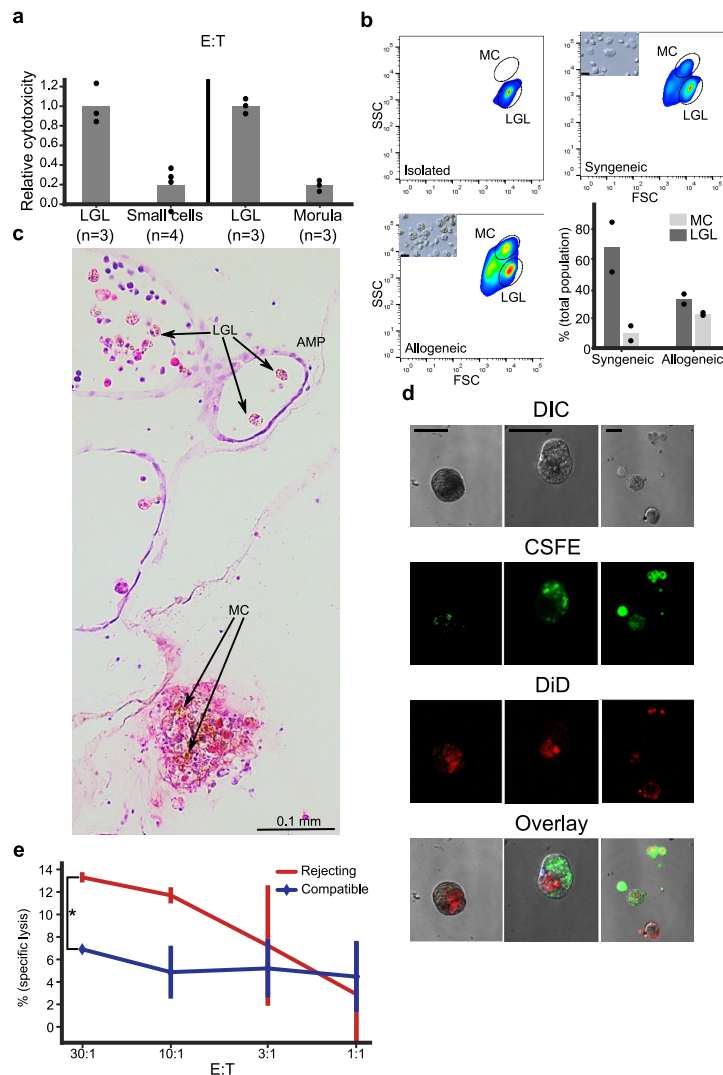


Extended Data Fig. 6 | Gene expression in an HSC niche: the endostyle. **a**, Comparison between the transcriptome sequence data from 10 samples of dissected endostyles and the transcriptome data for 34 whole colonies revealed a list of 327 genes that were significantly upregulated in the endostyle and showed homology to genes expressed in human haematopoietic bone marrow. Heat map includes the top 100 (by log₂(FC)) of the bone-marrow-associated endostyle genes. **b**, Geneset Activity Analysis of top 200 genes upregulated in the *B. schlosseri* endostyle

associated with the blood system, found by RNA-seq (this study) using the Gene Expression Commons tool on a mouse haematopoiesis model. The enriched populations are bone marrow stromal cells and HSCs. **c**, Similar analysis, but for *C. robusta* based on previous *in situ* work²¹, revealed enriched mouse bone marrow stromal cells as well, based on 188 genes that are expressed in the *C. robusta* endostyle and are associated with the blood system.



Extended Data Fig. 7 | Discovery of a myeloid lineage phagocytic population. **a**, FACS analysis of *B. schlosseri* cells that are fluorescently positive in one of three phagocytosis assays performed: (first and second from left) phagocytosis of green fluorescent beads, (third) phagocytosis of *V. diazotrophicus* labelled with AF647, and (fourth) allogeneic phagocytosis. Three phagocytic populations were identified: amoebocytes (A), myeloid cells (M), and large phagocytes (LP). The experiment was repeated twice. The myeloid cells were the main contributors to phagocytosis, contributing more than 40% to each of the phagocytosis assays. The large phagocytes contribute mainly to allogeneic phagocytosis compared to the other assays. **b**, Live images of the three isolated phagocytic populations. The experiment was performed three times. Scale bar, 20 μ m. **c**, We isolated the three main phagocytic populations and a small cell population (CP3) as a control, and incubated each one with fluorescent beads to validate the engulfment capacity of each population. The experiment was repeated twice. Plots show FACS analysis of green fluorescent bead phagocytosis by sorted populations. **d**, Amoebocytes, myeloid cells, and large phagocytes all had significantly higher engulfment rates than the small cell population. Moreover, amoebocytes and myeloid cells had significantly higher cell percentages than the large phagocyte population. Percentage analysis was carried out on two samples for each sorted population. Unpaired *t*-test, two-tailed; * $P < 0.05$, data shown as mean. **e**, Representative confocal images of the three phagocytic populations after engulfment of beads. Scale bar, 20 μ m. **f**, ImageStream analysis confirmed that the three phagocytic populations assayed engulfed the beads. The positive cells had mainly the morphology of amoebocytes, myeloid cells, or large phagocytes. The experiment was performed once on ImageStream. Representative images of the three phagocytic populations after engulfment of beads. Scale bar, 7 μ m.



Extended Data Fig. 8 | Cytotoxicity and the two morphs of morula cells at PORs. **a**, Cytotoxicity assays of isolated LGL cells compared to small cells, and to isolated morula cells (MCs). In both cases the LGL cells had significantly higher cytotoxicity compared to the other cell types. The experiment with isolated cells was performed twice with triplicates. Unpaired *t*-test, two-tailed; **P* = 0.003, ***P* = 0.0013; data shown as mean. **b**, LGL cells were isolated (upper left) and incubated overnight either in syngeneic (upper right) or in allogeneic challenge (lower left). FSC/SSC analysis of LGL cell (lower population) and morula cells (upper population). Insets, sample light microscopy images of the cells after incubation for each treatment. Lower right, analysis of LGL and morula cells in syngeneic compared to allogeneic challenge. The experiment was performed once with duplicates and validated by light microscopy. Bars show mean. **c**, H&E-stained section of *B. schlosseri* colonies undergoing

rejection. In the ampule (AMP) the inactivated form of cytotoxic morula cells/large granular lymphocyte-like cells (LGL) can be observed (top). On the other hand, the activated form with the brown pigmentation of morula cells can be observed at POR (bottom). **d**, Confocal imagery of phagocytosis assays to validate the allogeneic engulfment. Colonies are labelled with CFSE in green and with DiD in red after allogeneic phagocytosis assay. Large phagocytic cells can be seen after engulfment of allogeneic cells or vesicles. Validation of allogeneic phagocytosis by confocal imaging was performed twice. Scale bar, 20 μ m. **e**, Example of cytotoxicity assay with different effector to target (E:T) ratios, where the targets are compatible or rejecting colony cells to the effector colony. In the rejecting colony, specific lysis is significantly higher. The experiment was performed three times with triplicates. ANOVA two-factor with replication; **P* = 0.0015; data shown as mean \pm s.d.

Extended Data Table 1 | Antibodies screened by CyTOF for binding of *B. schlosseri* cells

| Symbol | Mass | Panel1 antigen | Clone | Supplier | CyTOF | FACS | Panel2 antigen | Clone | Supplier | CyTOF | FACS |
|--------|---------|----------------|-----------|--------------|-----------|------------|----------------|-----------|----------------------|-----------|------|
| Pd/Cd | 110-114 | CD3 | S4.1 | Invitrogen | Yes | No | | | | | |
| In | 113 | CD7 | M-T701 | BD | Yes | No | | | | | |
| In | 115 | CD45 | HI30 | Biolegend | Yes | No | | | | | |
| La | 139 | CD57 | HCD57 | Biolegend | Yes (low) | Yes (~14%) | CD2 | RPA-2.10 | Biolegend | Yes (low) | No |
| Pr | 141 | NKp46 | 195314 | R&D system | No | | CD61 | VI-PL2 | Biolegend | No | |
| Nd | 142 | CD48 | TU145 | BD | No | | | | | | |
| Nd | 144 | CCR5 | HEK/1/85a | Biolegend | Yes | N/A | CD94 | DX22 | Biolegend | Yes (low) | N/A |
| Nd | 145 | | | | | | LILRB1 | 292319 | R&D | No | |
| Nd | 146 | | | | | | CD309 | 89106 | BD | No | |
| Sm | 147 | | | | | | CD8 | RPA-T8 | Biolegend | No | |
| Nd | 148 | KIR3DL1 | DX9 | BD | No | | CRTAM | Cr24.1 | Biolegend | No | |
| Sm | 149 | DNAM1 | DX11 | BD | No | | | | | | |
| Eu | 151 | NKG2D | 1D11 | Biolegend | No | | CD84 | CD84.1.21 | Biolegend | No | |
| Sm | 152 | TNFR2 | 22235 | R&D | No | | | | | | |
| Eu | 153 | NKG2C | 134522 | R&D | Yes (low) | N/A | | | | | |
| Sm | 154 | NKp44 | P44-8 | Biolegend | No | | Notch1 | MHN1-519 | Biolegend | No | |
| Gd | 155 | CRACC | 162.1 | Biolegend | No | | | | | | |
| Gd | 156 | KIR2DL3 | 180701 | R&D | No | | | | | | |
| Gd | 158 | CD161 | HP-3G10 | Biolegend | No | | clec12A | 50C1 | Biolegend | No | |
| Tb | 159 | | | | | | CD11c | Bu15 | Biolegend | No | |
| Gd | 160 | NKp30 | 210845 | R&D | No | | 2B4 | C1.7 | Beckman | No | |
| Dy | 161 | CD15 | W6D3 | Biolegend | No | | CD4 | RPA-T4 | Biolegend | Yes | No |
| Dy | 162 | CD49d | 9F10 | Biolegend | Yes | Yes (~28%) | | | | | |
| Dy | 163 | CD16 | 3G8 | Biolegend | No | | KIR2DLS1 | EB6.B | Beckman | No | |
| Er | 166 | TIGIT | MBSA43 | eBioscience | No | | KIR3DL2 | DX31 | Gift from Dr. Lanier | No | |
| Er | 167 | CD27 | O323 | Biolegend | No | | CCR7 | G043H7 | Biolegend | No | |
| Er | 168 | KIR2DL1 | 143211 | R&D | No | | CCR2 | 48607 | BD | Yes | No |
| Er | 170 | CD11a | HI111 | Biolegend | No | | CD11b_act | CBRM1/5 | Biolegend | No | |
| Yb | 172 | KIR2DL4 | 181703 | R&D | No | | CD34 | 8G12 | BD | Yes (low) | N/A |
| Yb | 173 | | | | | | CD33 | WM53 | Biolegend | No | |
| Yb | 174 | CD144 | TEA1/31 | Beckman | No | | NKG2A | Z199 | Beckman | No | |
| Lu | 175 | KLRG1 | 13F12F2 | ThermoFisher | No | | | | | | |
| Yb | 176 | | | | | | CD56 | NCAM16.2 | BD | Yes (low) | N/A |
| Ir | 191/193 | DNA | | Sigma | All cells | | DNA | | Sigma | All cells | |
| Pt | 195 | cisplatin | | Sigma | No | | | | Sigma | | |

A screen of 49 antibodies that would cross-react with *B. schlosseri* cells and could potentially differentiate cellular populations. Symbol represents element; mass shows elemental isotope mass; antigen is the human antigen against which the antibody was produced. CyTOF column represents whether the *B. schlosseri* cell population was positive or not; low means less than 1% of cells were positive. FACS column represents whether there was binding by the same antibody clone by flow cytometry.

Extended Data Table 2 | Flow cytometry binding of *B. schlosseri* cells, and references for phagocytic and cytotoxic cells in different organisms

| | | |
|--|--|--|
| a | | |
| Lectin | Fluorophore | % positive cells |
| ConA | AF-633 | 45% |
| PNA | PE-Cy7 | 16% |
| UEA | PE-Cy7 | 30% |
| Additional markers | | |
| Alkaline Phosphatase | Green | 9% high 25% mid |
| Anti-BHF polyclonal | AF-647 | 35% |
| b | | |
| Animal: | Phagocytic cells | Cytotoxic cells |
| <i>Homo</i> (human) | Myeloid lineage review ⁵² | Review ⁵³ |
| <i>Danio</i> (zebrafish) | Myeloid lineage review ⁵⁴ , phagocytosis ⁵⁵ | Recognition molecules ⁵⁶ , cytotoxicity ⁵⁵ , cell types ⁵⁷ |
| <i>Petromyzon</i> (lamprey) | Myeloid lineage ⁵⁸ | Lymphoid lineage review ⁵⁹ (cytotoxicity was not shown) |
| <i>Branchiostoma</i> (lancelet) | Large phagocytes ⁶⁰ , some suggestion of amoebocytes ⁶¹ | |
| <i>Strongylocentrotus</i> (sea urchin) | Phagocytic cells review ³² | Suggestion that colorless spherule cells are cytotoxic morula cells in review ³² |
| <i>Drosophila</i> (fruit fly) | Amoebocytes ⁶² , plasmatocytes or pupal macrophage-like have resemblance to the large phagocytes ^{63,64} | Crystal cells that contain the enzymes for melanization for cytotoxicity could resemble the morula cells at the enzymatic level ^{63,64} |
| <i>Limulus</i> (horseshoe crab) | Amoebocytes and granular phagocytes in review ³³ | Cells with prophenoloxidase and melanization process in review ^{33,34} |
| <i>Tridacna</i> (clam) | Amoebocytes and eosinophilic granular hemocytes that resemble large phagocytes ⁶⁵ | Morula cells ⁶⁵ |

a. The column '% positive cells' shows the percentage of the cells that were positively labelled by the marker. For alkaline phosphatase (AP), 'high' represents cells that labelled strongly and 'mid' represents cells that were positively but not strongly labelled. **b.** Rows contain references^{52–65} and any notes for each of the organisms reviewed in the production of Fig. 4.

Consistent success in life-supporting porcine cardiac xenotransplantation

Matthias Längin^{1,2,18}, Tanja Mayr^{1,2,18}, Bruno Reichart^{2*}, Sebastian Michel³, Stefan Buchholz³, Sonja Guethoff^{2,3}, Alexey Dashkevich³, Andrea Baehr⁴, Stefanie Egerer⁴, Andreas Bauer¹, Maks Mihalj³, Alessandro Panelli², Lara Issl², Jiawei Ying², Ann Kathrin Fresch², Ines Buttgereit², Maren Mokelke², Julia Radan², Fabian Werner¹, Isabelle Lutzmann², Stig Steen⁵, Trygve Sjöberg⁵, Audrius Paskevicius⁵, Liao Qiuming⁵, Riccardo Sfriso⁶, Robert Rieben⁶, Maik Dahlhoff⁴, Barbara Kessler⁴, Elisabeth Kemter⁴, Katharina Klett^{7,8,9}, Rabea Hinkel^{7,8,9}, Christian Kupatt^{7,9}, Almuth Falkenau¹⁰, Simone Reu¹¹, Reinhard Ellgass³, Rudolf Herzog³, Uli Binder¹², Günter Wich¹³, Arne Skerra¹⁴, David Ayares¹⁵, Alexander Kind¹⁶, Uwe Schönmann¹⁷, Franz-Josef Kaup¹⁷, Christian Hagl¹³, Eckhard Wolf⁴, Nikolai Klymiuk⁴, Paolo Brenner^{2,3,19} & Jan-Michael Abicht^{1,2,19}

Heart transplantation is the only cure for patients with terminal cardiac failure, but the supply of allogeneic donor organs falls far short of the clinical need^{1–3}. Xenotransplantation of genetically modified pig hearts has been discussed as a potential alternative⁴. Genetically multi-modified pig hearts that lack galactose- α 1,3-galactose epitopes (α 1,3-galactosyltransferase knockout) and express a human membrane cofactor protein (CD46) and human thrombomodulin have survived for up to 945 days after heterotopic abdominal transplantation in baboons⁵. This model demonstrated long-term acceptance of discordant xenografts with safe immunosuppression but did not predict their life-supporting function. Despite 25 years of extensive research, the maximum survival of a baboon after heart replacement with a porcine xenograft was only 57 days and this was achieved, to our knowledge, only once⁶. Here we show that α 1,3-galactosyltransferase-knockout pig hearts that express human CD46 and thrombomodulin require non-ischæmic preservation with continuous perfusion and control of post-transplantation growth to ensure long-term orthotopic function of the xenograft in baboons, the most stringent preclinical xenotransplantation model. Consistent life-supporting function of xenografted hearts for up to 195 days is a milestone on the way to clinical cardiac xenotransplantation⁷.

Xenotransplantation of genetically multi-modified α 1,3-galactosyltransferase-knockout pig hearts that express human CD46 and thrombomodulin (blood group 0) was performed using the clinically approved Shumway's orthotopic technique⁸. Fourteen captive-bred baboons (*Papio anubis*, blood groups B and AB) served as recipients. All recipients received basic immunosuppression, similar to that described previously⁵: induction therapy included anti-CD20 antibody, anti-thymocyte-globulin, and the monkey-specific anti-CD40 mouse/rhesus chimeric IgG4 monoclonal antibody (clone 2C10R4)⁹ or our own humanized anti-CD40L PASylated (conjugated with a long, structurally disordered Pro-Ala-Ser amino acid chain) antigen-binding fragment (Fab)¹⁰. During maintenance therapy methylprednisolone was reduced gradually, whereas mycophenolate mofetil and anti-CD40 monoclonal antibody or anti-CD40L PASylated Fab treatment remained constant (Extended Data Table 1). Postoperative treatment of the recipients has been described elsewhere¹¹.

In group I ($n = 5$), donor organs were preserved with two clinically approved crystalloid solutions (4 °C custodiol HTK

(histidine-tryptophan-ketoglutarate) or Belzer's UW solution), each perfused after cross-clamping the ascending aorta before excision of the porcine donor organ. The hearts were kept in plastic bags filled with ice-cold solution and surrounded by ice cubes (static preservation).

The results of group I were disappointing. Despite short ischaemic preservation periods (123 ± 7 min), the animals survived for only 1 day ($n = 3$), 3 days ($n = 1$) and 30 days ($n = 1$) (Fig. 1a). The four short-term survivors were successfully taken off cardiopulmonary bypass (CPB) and three could be extubated, but all were lost due to severe systolic left heart failure in spite of a high dose of intravenous catecholamines (Extended Data Fig. 1). This so-called 'perioperative cardiac xenograft dysfunction' (PCXD)¹² has been observed in 40 to 60% of the orthotopic cardiac xenotransplantation experiments described in the literature⁴. The only 30-day survivor (which received a heart preserved with Belzer's UW solution) gradually developed left ventricular myocardial hypertrophy and stiffening, resulting in progressive diastolic left ventricular failure associated with increased serum levels of troponin T, an indicator of myocardial damage (Fig. 1b). Increased serum bilirubin levels (Fig. 1c) and several other clinically relevant chemical parameters (Table 1) indicated associated terminal liver disease. Upon necropsy, marked cardiac hypertrophy (Fig. 1e) with a thickened left ventricular myocardium and a decreased left ventricular cavity was evident (Fig. 1f).

To reduce the incidence of the PCXD that was observed in group I, we explored new ways to improve xenograft preservation. In group II ($n = 4$), the same immunosuppressive regime as in group I was used, but the pig hearts were preserved with an 8 °C oxygenated albumin-containing hyperoncotic cardioplegic solution that contained nutrition, hormones and erythrocytes¹³. From explantation until transplantation, the organs were continuously perfused and oxygenated using a heart-perfusion system. During implantation surgery, the hearts were intermittently perfused every 15 min until the aortic clamp was opened at the end of transplantation.

After non-ischæmic continuous organ perfusion (206 ± 43 min), all four baboons in group II could easily be taken off CPB, showed better graft function compared to animals in group I and required less catecholamine support (Extended Data Fig. 1). No organ was lost owing to PCXD. One experiment had to be terminated on the fourth postoperative day because of a technical failure; the other three animals lived for 18, 27 and 40 days (Fig. 1a). Echocardiography during the experiments revealed increasing hypertrophy of the left ventricular

¹Department of Anaesthesiology, University Hospital, LMU Munich, Munich, Germany. ²Transregional Collaborative Research Center 127, Walter Brendel Centre of Experimental Medicine, LMU Munich, Munich, Germany. ³Department of Cardiac Surgery, University Hospital, LMU Munich, Munich, Germany. ⁴Institute of Molecular Animal Breeding and Biotechnology, Gene Center, LMU Munich, Munich, Germany. ⁵Department of Cardiothoracic Surgery, Lund University and Skåne University Hospital, Lund, Sweden. ⁶Department for BioMedical Research (DMBR), University of Bern, Bern, Switzerland. ⁷I. Medizinische Klinik, Klinikum Rechts der Isar, Technical University of Munich, Munich, Germany. ⁸Institute for Cardiovascular Prevention (IPEK), LMU Munich, Munich, Germany. ⁹DZHK (German Center for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany. ¹⁰Institute of Veterinary Pathology, LMU Munich, Munich, Germany. ¹¹Institute of Pathology, Medical Faculty, LMU Munich, Munich, Germany. ¹²XL-protein GmbH, Freising, Germany. ¹³Wacker-Chemie AG, Munich, Germany. ¹⁴Munich Center for Integrated Protein Science (CIPS-M) and Lehrstuhl für Biologische Chemie, School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany. ¹⁵Revivicor, Blacksburg, VA, USA. ¹⁶Chair of Livestock Biotechnology, School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany. ¹⁷German Primate Centre, Göttingen, Germany. ¹⁸These authors contributed equally: Matthias Längin, Tanja Mayr. ¹⁹These authors jointly supervised this work: Paolo Brenner, Jan-Michael Abicht. *e-mail: bruno.reichart@med.uni-muenchen.de

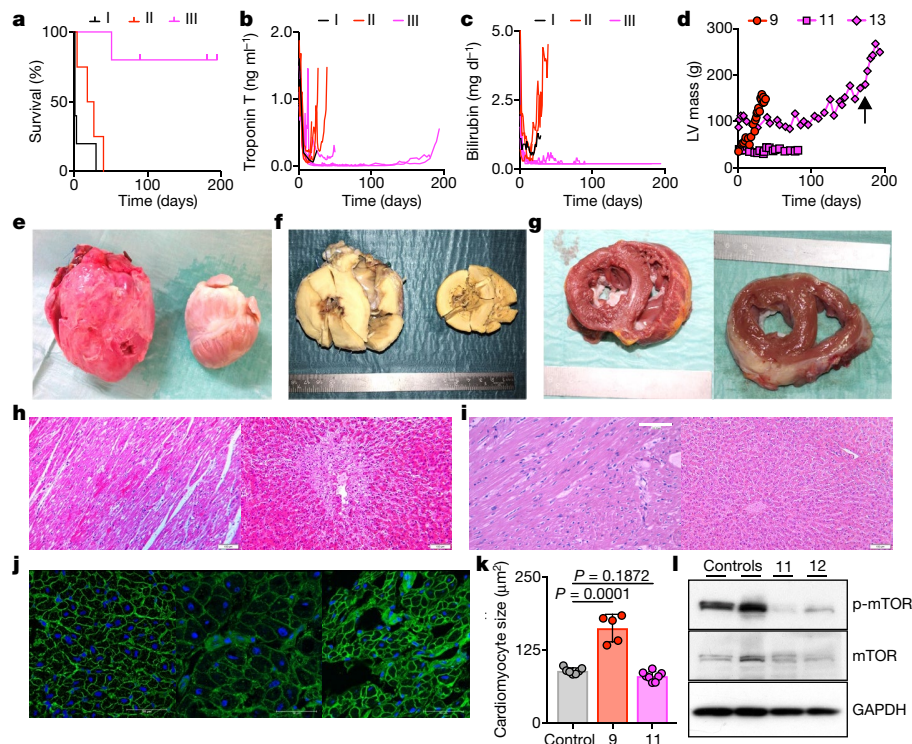


Fig. 1 | Survival, laboratory parameters, necropsy and histology after orthotopic xenotransplantation. **a**, Kaplan–Meier curve of survival of groups I (black; $n = 5$ animals), II (red; $n = 4$ animals) and III (magenta; $n = 5$ animals). Two-sided log-rank test, $P = 0.0007$. **b**, **c**, Serum concentrations of cardiac troponin T (**b**) and bilirubin (**c**). **d**, Left ventricular (LV) masses of xenografted hearts from animals 9 (group II), 11 and 13 (both group III); note increased graft growth after discontinuation of temsirolimus (arrow). **e–g**, Front view of the porcine donor heart and own heart of baboon 3 (**e**, left and right, respectively; group I) and transverse cuts of the porcine donor hearts (left) and the baboons' own hearts (right) of animals 3 (**f**) and 11 (**g**). Note the extensive left ventricular hypertrophy and reduction of left ventricular cavity of the donor organ of baboon 3 in contrast to animal 11. **h**, **i**, Haematoxylin and eosin staining of the left ventricular myocardium of the donor (left) and the liver of the recipient (right). Scale bars, 100 μm . **h**, The myocardium of animal 9 showed multifocal cell necroses with hyper eosinophilia, small vessel thromboses, moderate interstitial infiltration of lymphocytes,

neutrophils and macrophages. The liver of this animal had multifocal centrilobular cell vacuolizations and necroses as well as multifocal intralesional haemorrhages. **i**, The myocardium of baboon 11 had sporadic infiltrations of lymphocytes, multifocal minor interstitial oedema whereas the liver had small vacuolar degeneration of hepatocytes (lipid type). **j**, Wheat germ agglutinin-stained myocardial sections of a sham-operated porcine heart (left), and the hearts transplanted into animals 9 (centre) and 11 (right). Scale bar, 50 μm . **e–j**, $n = 4$, groups I/II; $n = 3$, group III; $n = 1$, control; one representative biological sample for each group is shown for group I/II, group III and control (**j**). **k**, Quantitative analysis of cardiomyocyte cross-sectional areas. Data are mean \pm s.d., P values are indicated, one-way analysis of variance (ANOVA) with Holm–Sidak's multiple comparisons test ($n = 3$ biologically independent samples with 5–8 measurements each). **l**, Western blot analysis of myocardium from transplanted hearts of animals 11 and 12 showed reduced mTOR phosphorylation (p-mTOR) compared to age-matched control samples. $n = 2$, group III; $n = 2$, controls. For gel source data, see Supplementary Fig. 1.

myocardium as measured by left ventricular mass^{14,15} (Fig. 1d), left ventricular stiffening and decreasing left ventricular filling volumes (Extended Data Fig. 2a). Graft function remained normal throughout the experiments, but diastolic relaxation gradually deteriorated (Supplementary Video 1). Troponin T levels were consistently above normal range and increased markedly at the end of each experiment (Table 1 and Fig. 1b) and simultaneously platelet counts decreased whereas lactate dehydrogenase (LDH) increased (Table 1 and Extended Data Fig. 3a, b), suggesting thrombotic microangiopathy as described for heterotopic abdominal cardiac xenotransplantation^{5,16}. In addition, secondary liver failure developed: increasing serum bilirubin concentrations (Fig. 1c) and a decrease in prothrombin ratio and reduction in cholinesterase indicated a reduction in liver function, while increased serum activities of alanine aminotransferase (ALT) and aspartate aminotransferase (AST) pointed to liver damage (Table 1). At necropsy, the weight of group II hearts had more than doubled (on average 259%) compared to the time point of transplantation. Histology confirmed myocardial cell hypertrophy (Fig. 1j, k) and revealed multifocal myocardial necroses, thromboses and immune cell infiltration (Fig. 1h); in the liver, multifocal cell necroses were observed (Fig. 1h). Taken together, these alterations are consistent with diastolic pump failure and subsequent congestive liver damage resulting from massive cardiac overgrowth. However, immunofluorescence analyses of the myocardium and plasma levels of non-galactose- α 1,3-galactose xenoreactive

antibodies¹⁷ did not indicate humoral rejection of the graft (Fig. 2 and Extended Data Fig. 4).

To prevent diastolic heart failure, we investigated means of reducing cardiac hypertrophy. The following modifications were made for group III ($n = 5$): recipients were weaned from cortisone at an early stage and received antihypertensive treatment (pigs have a lower systolic blood pressure than baboons, around 80 compared to approximately 120 mm Hg, respectively) and additional temsirolimus medication was used to counteract cardiac overgrowth. After heart perfusion times of 219 ± 30 min, all five animals were easily taken off CPB, comparable to group II (Extended Data Fig. 1). None of the recipients in group III showed PCXD; all reached a steady state with good heart function after four weeks. One recipient (10) developed recalcitrant pleural effusions that were caused by occlusion of the thoracic lymph duct and was therefore euthanized after 51 days. Two recipients (11, 12) lived in good health for three months until euthanasia, according to the study protocol (Fig. 1a). In these three recipients, echocardiography revealed no increase in left ventricular mass (Fig. 1d); graft function remained normal with no signs of diastolic dysfunction (Extended Data Fig. 2b and Supplementary Video 2). Biochemical parameters of heart and liver functions as well as LDH levels and platelet counts were normal or only slightly altered throughout the experiments (Table 1, Fig. 1b, c and Extended Data Fig. 3a, b), consistent with normal histology (Fig. 1i). Histology of left ventricular myocardium showed no

Table 1 | Serum levels of liver and heart enzymes, platelet counts and prothrombin ratio at the end of experiments that lasted longer than two weeks

| | Group I | Group II | Group III | | | | | | | |
|---|-------------------------|-------------------------|-------------------------|-------------------------|---|------------|------------|------------|------------|-----------|
| Experiment | 3 | 6 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | Reference |
| Bilirubin (mg dl ⁻¹) | 1.2 | 0.9 | 2.7 | 4.5 | 0.3 | 0.2 | 0.2 | 0.2 | 0.2 | ≤1.2 |
| AST (U l ⁻¹) | 646 | 896 | 792 | 354 | 101 | 27 | 23 | 63 | 28 | ≤49 |
| PR (%) | 30 | 6 | 6 | 6 | 101 | 96 | 117 | 26 | 99 | 70–130 |
| CHE (kU l ⁻¹) | 1.6 | 1.6 | 1.4 | 1.1 | 2.1 | 9.4 | 14.4 | 7.3 | 7.2 | 4.6–11.5 |
| Troponin T (ng ml ⁻¹) | 0.233 | 0.660 | 1.460 | 1.470 | 0.218 | 0.037 | 0.018 | 0.556 | 0.140 | ≤0.014 |
| CK total (U l ⁻¹) | 654 | 636 | 1017 | 953 | 3053 | 143 | 66 | 461 | 96 | ≤189 |
| LDH (U l ⁻¹) | 3252 | 6853 | 2842 | 1627 | 436 | 311 | 511 | 962 | 497 | ≤249 |
| Platelets (billion particles per litre) | 99 | 101 | 65 | 29 | 216 | 202 | 128 | 271 | 303 | 150–300 |
| Survival (days) | 30 | 18 | 27 | 40 | 51 | 90 | 90 | 195 | 182 | |
| Causes of death | Heart and liver failure | Heart and liver failure | Heart and liver failure | Heart and liver failure | SVC thrombosis, thoracic duct occlusion | Euthanasia | Euthanasia | Euthanasia | Euthanasia | |

Normal reference values are given in the right-most column. Animals from groups I and II exhibited pathological biochemical alterations that correspond to heart and liver failure; platelet counts were low and LDH was elevated. By contrast, most parameters remained close to, or within, normal ranges in animals of group III. The baboon in experiment 10 had to be euthanized because of severe pleural effusions due to superior vena cava (SVC) thrombosis and occlusion of the thoracic lymph duct. The animals in experiments 11 and 12 were euthanized after reaching the study end point of 90 days, although they did not show any signs of cardiac or liver dysfunction. Animals in experiments 13 and 14 were euthanized after six months; recipient 13 showed the signs of beginning heart and liver dysfunction. CHE, cholinesterase; CK, creatine kinase; PR, prothrombin ratio.

signs of hypertrophy (Fig. 1j, k), and western blot analysis of the myocardium revealed phosphorylation levels of mTOR that were lower than non-transplanted age-matched control hearts (Fig. 1l). Similar to group II, there were no signs of humoral graft rejection in group III (Fig. 2 and Extended Data Fig. 4).

The study protocol for group III was extended aiming at a graft survival of six months. The last two recipients in this group (13, 14) were allowed to survive in good general condition for 195 and 182 days, with no major changes to platelet counts or serum LDH and bilirubin levels (Fig. 1a, c and Extended Data Fig. 3a, b). Intravenous temsirolimus treatment was discontinued on day 175 and on day 161. Up to this point, systolic and diastolic heart function was normal (Supplementary Video 3). Thereafter, increased growth of the cardiac graft was observed in both recipients (Fig. 1d), emphasizing the importance of mTOR inhibition in the orthotopic xenogeneic heart xenotransplantation model.

Similar to the changes observed in group II, the smaller recipient 13 developed signs of diastolic dysfunction, which was associated with elevated serum levels of troponin T and the start of congestive liver damage (increased serum ALT and AST levels, decreased prothrombin ratio and cholinesterase); platelet counts remained within normal ranges (Table 1, Fig. 1b, c and Extended Data Fig. 3a, b). Histology confirmed hepatic congestion and revealed multifocal myocardial necroses without immune cell infiltrations or signs of thrombotic microangiopathy. In the larger recipient 14, who had to be euthanized simultaneously with animal 13, the consequences of cardiac overgrowth were minimal.

Here we show consistent survival of life-supporting pig hearts in non-human primates for at least three months that meets the preclinical efficacy requirements for the initiation of clinical xenotransplantation trials as suggested by an advisory report of the International Society for Heart and Lung Transplantation⁷.

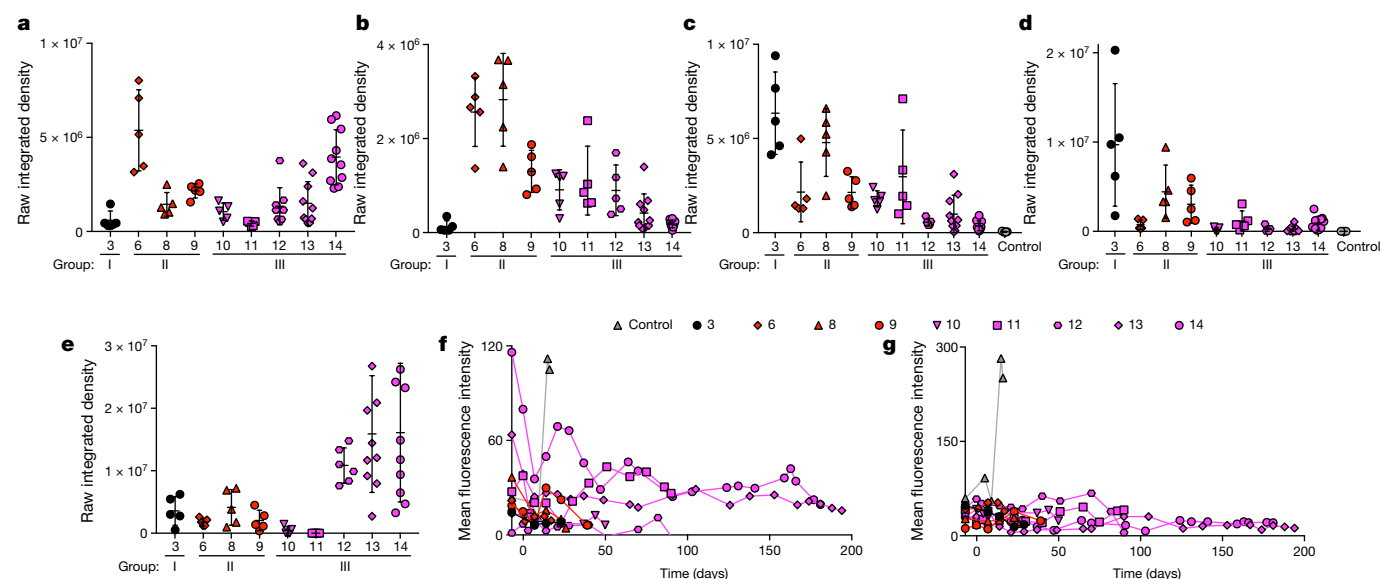


Fig. 2 | Quantitative evaluation of antibodies, complement and fibrin in myocardial tissue and serum levels of non-galactose-α1,3-galactose xenoreactive antibodies. a–e, Quantitative evaluation of fluorescence intensities ($n = 9$ biologically independent samples with 5–10 measurements per experiment; for representative images see Extended Data Fig. 4). Raw integrated densities are shown for IgM (a), IgG (b), C3b/c (c), C4b/c (d) and fibrin (e). Group I (animal 3), black; group II (animals 6, 8, 9), red; group III (animals 11–14), magenta. C3b/c and C4b/c values are compared to those of controls measured in healthy pig

hearts. Data are mean \pm s.d. f, g, Levels of non-galactose-α1,3-galactose xenoreactive IgM and IgG antibodies in baboon plasma; antibody binding to α-galactosyltransferase-knockout porcine aortic endothelial cells that express human CD46 and thrombomodulin was analysed by fluorescence-activated cell sorting. Values are expressed as mean fluorescence intensity. Animals 6, 9 and 10 received an anti-CD40L PASylated Fab, the others were treated with an anti-CD40 monoclonal antibody. Plasma from a baboon who rejected a heterotopically intrathoracic transplanted pig heart served as positive control (grey).

Two steps were key to success. First, non-ischaemic porcine heart preservation was found to be important for the survival of the xenografted hearts. Xenografted hearts from group I that underwent ischaemic static myocardial preservation with crystalloid solutions (as used for clinical allogeneic procedures) showed PCXD in four out of five cases, necessitating higher amounts of catecholamines. This phenomenon is clearly similar to 'cardiac stunning', the occurrence of which has been known since the early days of cardiac surgery and does not represent hyperacute rejection⁴. By contrast, in groups II and III (non-ischaemic porcine heart preservation by perfusion)¹³, all nine recipients came off CPB easily since their cardiac outputs remained unchanged compared to baseline. The short-term results achieved in these groups were excellent even by clinical standards.

The second key step was the prevention of detrimental xenograft overgrowth. Previous pig-to-baboon kidney and lung transplantation experiments have suggested that growth of the graft depends more on intrinsic factors than on stimuli from the recipient such as growth hormones¹⁸. The massive cardiac hypertrophy in our group-II recipients indicates a more complex situation. Notably, a transplanted heart in this group had a 62% greater weight gain than the non-transplanted heart of a sibling in the same time span (Extended Data Fig. 2c).

In group III, cardiac overgrowth was successfully counteracted by a combination of treatments: (i) decreasing the blood pressure of the baboons to match the lower porcine levels; (ii) tapering cortisone at an early stage—cortisone can cause hypertrophic cardiomyopathy in early life in humans¹⁹; and (iii) using the sirolimus prodrug temsirolimus to mitigate myocardial hypertrophy. Sirolimus compounds are known to control the complex network of cell growth by inhibiting both mTOR kinases²⁰. There is clinical evidence that sirolimus treatment can attenuate myocardial hypertrophy and improve diastolic pump function^{21,22}, as well as ameliorate rare genetic overgrowth syndromes in humans²³. In addition to the effects of human thrombomodulin expression in the graft^{5,24}, temsirolimus treatment may prevent the formation of thrombotic microangiopathic lesions even further by reducing collagen-induced platelet aggregation and by destabilizing platelet aggregates formed under shear stress conditions²⁵.

In summary, our study demonstrates that consistent long-term life-supporting orthotopic xenogeneic heart transplantation in the most relevant preclinical model is feasible, facilitating clinical translation of xenogeneic heart transplantation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0765-z>.

Received: 18 April 2018; Accepted: 2 November 2018;

Published online 5 December 2018.

- Lund, L. H. et al. The Registry of the International Society for Heart and Lung Transplantation: thirty-fourth adult heart transplantation report—2017; focus theme: allograft ischemic time. *J. Heart Lung Transplant.* **36**, 1037–1046 (2017).
- Rossano, J. W. et al. The Registry of the International Society for Heart and Lung Transplantation: twentieth pediatric heart transplantation report—2017; focus theme: allograft ischemic time. *J. Heart Lung Transplant.* **36**, 1060–1069 (2017).
- Eurotransplant. *Annual Report 2016*. **103** (Eurotransplant International Foundation, 2017).
- Mohiuddin, M. M., Reichart, B., Byrne, G. W. & McGregor, C. G. A. Current status of pig heart xenotransplantation. *Int. J. Surg.* **23**, 234–239 (2015).
- Mohiuddin, M. M. et al. Chimeric 2C10R4 anti-CD40 antibody therapy is critical for long-term survival of GTKO.hCD46.hTBM pig-to-primate cardiac xenograft. *Nat. Commun.* **7**, 11138 (2016).
- Byrne, G. W., Du, Z., Sun, Z., Asmann, Y. W. & McGregor, C. G. Changes in cardiac gene expression after pig-to-primate orthotopic xenotransplantation. *Xenotransplantation* **18**, 14–27 (2011).
- Cooper, D. K. et al. Report of the Xenotransplantation Advisory Committee of the International Society for Heart and Lung Transplantation: the present status of xenotransplantation and its potential role in the treatment of end-stage cardiac and pulmonary diseases. *J. Heart Lung Transplant.* **19**, 1125–1165 (2000).
- Lower, R. R. & Shumway, N. E. Studies on orthotopic homotransplantation of the canine heart. *Surg. Forum* **11**, 18–19 (1960).
- Lowe, M. et al. A novel monoclonal antibody to CD40 prolongs islet allograft survival. *Am. J. Transplant.* **12**, 2079–2087 (2012).

- Binder, U. & Skerra, A. PASylation®: A versatile technology to extend drug delivery. *Curr. Opin. Colloid Interface Sci.* **31**, 10–17 (2017).
- Mayr, T. et al. Hemodynamic and perioperative management in two different preclinical pig-to-baboon cardiac xenotransplantation models. *Xenotransplantation* **24**, e12295 (2017).
- Byrne, G. W. & McGregor, C. G. Cardiac xenotransplantation: progress and challenges. *Curr. Opin. Organ Transplant.* **17**, 148–154 (2012).
- Steen, S., Paskevicius, A., Liao, Q. & Sjöberg, T. Safe orthotopic transplantation of hearts harvested 24 hours after brain death and preserved for 24 hours. *Scand. Cardiovasc. J.* **50**, 193–200 (2016).
- Devereux, R. B. et al. Echocardiographic assessment of left ventricular hypertrophy: comparison to necropsy findings. *Am. J. Cardiol.* **57**, 450–458 (1986).
- Lang, R. M. et al. Recommendations for chamber quantification: a report from the American Society of Echocardiography's Guidelines and Standards Committee and the Chamber Quantification Writing Group, developed in conjunction with the European Association of Echocardiography, a branch of the European Society of Cardiology. *J. Am. Soc. Echocardiogr.* **18**, 1440–1463 (2005).
- Kuwaki, K. et al. Heart transplantation in baboons using $\alpha 1,3$ -galactosyltransferase gene-knockout pigs as donors: initial experience. *Nat. Med.* **11**, 29–31 (2005).
- Azizmzadeh, A. M. et al. Development of a consensus protocol to quantify primate anti-non-Gal xenoreactive antibodies using pig aortic endothelial cells. *Xenotransplantation* **21**, 555–566 (2014).
- Tanabe, T. et al. Role of intrinsic (graft) versus extrinsic (host) factors in the growth of transplanted organs following allogeneic and xenogeneic transplantation. *Am. J. Transplant.* **17**, 1778–1790 (2017).
- Lesnik, J. J., Singh, G. K., Balfour, I. C. & Wall, D. A. Steroid-induced hypertrophic cardiomyopathy following stem cell transplantation in a neonate: a case report. *Bone Marrow Transplant.* **27**, 1105–1108 (2001).
- Saxton, R. A. & Sabatini, D. M. mTOR signaling in growth, metabolism, and disease. *Cell* **168**, 960–976 (2017).
- Imamura, T. et al. Everolimus attenuates myocardial hypertrophy and improves diastolic function in heart transplant recipients. *Int. Heart J.* **57**, 204–210 (2016).
- Paoletti, E. mTOR inhibition and cardiovascular diseases: cardiac hypertrophy. *Transplantation* **102**, S41–S43 (2018).
- Manning, B. D. Game of TOR — the target of rapamycin rules four kingdoms. *N. Engl. J. Med.* **377**, 1297–1299 (2017).
- Wuensch, A. et al. Regulatory sequences of the porcine THBD gene facilitate endothelial-specific expression of bioactive human thrombomodulin in single- and multitransgenic pigs. *Transplantation* **97**, 138–147 (2014).
- Aslan, J. E., Tormoen, G. W., Loren, C. P., Pang, J. & McCarty, O. J. S6K1 and mTOR regulate Rac1-driven platelet activation and aggregation. *Blood* **118**, 3129–3136 (2011).

Acknowledgements We thank the Walter Brendel Centre of Experimental Medicine, Munich for support and provision of facilities, especially U. Pohl, M. Shakarami and all animal caretakers. Financial support was provided by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) TRR 127. We acknowledge K. Reimann for providing the CD40 monoclonal antibody for the experiments.

Reviewer information Nature thanks C. Knosalla and J. Madsen for their contribution to the peer review of this work.

Author contributions B.R., P.B., T.M., M.L. and J.-M.A. conceived and led the study; M.L., T.M., B.R., R.E., R. Herzog, S.G., S.B., S.M., A.D., A. Bauer, F.W., M. Mihaj, A. Panelli, L.L., J.Y., A.K.F., L.Q., C.H., I.L., I.B., M. Mokelke, J.R., P.B. and J.-M.A. performed the experiments and collected samples; S.S., T.S., A. Paskevicius and L.Q. performed non-ischaemic heart preservation experiments; M.L., A. Panelli, A. Bauer and J.-M.A. performed haemodynamic and echocardiographic analyses; R.S. and R.R. provided immunological analyses; E.K., K.K., R. Hinkel and C.K. performed histochemical analyses; M.D. and E.W. provided protein analysis; A.F. and S.R. performed necropsy and histological analyses with contributions from T.M. and A. Panelli; M.L., T.M., B.R., R.S., R.R., R. Hinkel, M.D. and J.-M.A. analysed the data; A. Baehr, S.E., B.K., D.A., E.W. and N.K. provided genetically multi-modified donor pigs; U.S. and F.-J.K. provided non-human primates; U.B., G.W. and A.S. developed PASylated anti-CD40L Fab; B.R., M.L., T.M. and J.-M.A. wrote the manuscript; A.K., A.S., R.R., S.S., R. Hinkel, P.B. and E.W. reviewed and edited the manuscript.

Competing interests D.A. is chief executive officer and chief scientific officer of Revivicor, Inc. A.S. and U.B. are cofounders of XL-protein GmbH, Germany. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0765-z>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0765-z>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to B.R.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Animals. Experiments were carried out between February 2015 and August 2018. Fourteen juvenile pigs of cross-bred genetic background (German Landrace and Large White, blood group 0) served as donors for heart xenotransplantation. All organs were homozygous for $\alpha 1,3$ -galactosyltransferase knockout (GTKO), and heterozygous transgenic for human CD46 (hCD46) and human thrombomodulin (hTM)²⁴ (Revivicor and Institute of Molecular Animal Breeding and Biotechnology). Localization and stability of hCD46 and hTM expression were verified post mortem by immunohistochemistry (Extended Data Fig. 5). Donor heart function and absence of valvular defects were evaluated seven days before transplantation by echocardiography. Fourteen male captive-bred baboons (*P. anubis*, blood groups B and AB) were used as recipients (German Primate Centre).

The study was approved by the local authorities and the Government of Upper Bavaria. All animals were treated in compliance with the Guide for the Care and Use of Laboratory Animals (US National Institutes of Health and German Legislation).

Anaesthesia and analgesia. Baboons were premedicated by intramuscular injection of ketamine hydrochloride 6–8 mg kg⁻¹ (ketavet 100 mg ml⁻¹; Pfizer) and 0.3–0.5 mg kg⁻¹ midazolam (midazolam-ratiopharm; Ratiopharm). General anaesthesia was induced with an intravenous bolus of 2.0–2.5 mg kg⁻¹ propofol (propofol-lipuro 2%; B. Braun Melsungen) and 0.05 mg fentanyl (fentanyl-janssen 0.5 mg; Janssen-Cilag), and maintained with propofol (0.16 ± 0.06 mg kg⁻¹ min⁻¹) or sevoflurane (1–2 vol% endexpiratory; sevoflurane, AbbVie) and bolus administrations of fentanyl (6–8 µg kg⁻¹, repeated every 45 min) as described elsewhere¹¹. Continuous infusion of fentanyl, ketamine hydrochloride and metamizole (novaminsulfon-ratiopharm 1 g per 2 ml; Ratiopharm) was applied postoperatively to ensure analgesia.

Explantation and preservation of donor hearts. Pigs were premedicated by intramuscular injection of ketamine hydrochloride 10–20 mg kg⁻¹, azaperone 10 mg kg⁻¹ (stresnil 40 mg ml⁻¹; Lilly Deutschland) and atropine sulfate (atropin-sulfat B. Braun 0.5 mg; B. Braun Melsungen). General anaesthesia was induced with an intravenous bolus of 20 mg propofol and 0.05 mg fentanyl and maintained with propofol (0.12 mg kg⁻¹ min⁻¹) and bolus administrations of fentanyl (2.5 µg kg⁻¹, repeated every 30 min).

After median sternotomy and heparinization (500 IU kg⁻¹), a small cannula was inserted into the ascending aorta, which was then cross-clamped distal of the cannula. In group I, the heart was perfused with a single dose of 20 ml kg⁻¹ crystalloid cardioplegic solution at 4°C: custodiol HTK solution (Dr. Franz Köhler Chemie) was used for the hearts for animals 2, 4 and 5, Belzer's UW solution (Preservation Solutions) was used for the hearts for animals 1 and 3. The appendices of the right and left atrium were opened for decompression. The heart was then excised, submersed in cardioplegic solution and stored on ice.

In groups II and III, hearts were preserved as described previously¹³, using 3.5 l of an oxygenated albumin-containing hyperoncotic cardioplegic nutrition solution with hormones and erythrocytes at a temperature of 8°C in a portable extracorporeal heart preservation system consisting of a pressure- and flow-controlled roller pump, an O₂/CO₂ exchanger, a leukocyte filter, an arterial filter and a cooler/heater unit.

After aortic cross-clamping, the heart was perfused with 600 ml preservation medium, excised and moved into the cardiac preservation system. A large cannula was introduced into the ascending aorta and the mitral valve was made temporarily incompetent to prevent left ventricular dilation; the superior vena cava was ligated; however, the inferior vena cava, pulmonary artery and pulmonary veins were left open for free outlet of perfusate. The heart was submersed in a reservoir filled with cold perfusion medium and antegrade coronary perfusion commenced through the already placed aortic cannula. The perfusion pressure was regulated at exactly 20 mm Hg. During implantation, the heart was intermittently perfused for 2 min every 15 min.

Implantation technique. The recipient's thorax was opened at the midline. Unfractionated heparin (500 IU kg⁻¹; heparin-natrium-25000-ratiopharm, Ratiopharm) was given and the heart–lung machine connected, using both venae cavae and the ascending aorta. CBP commenced and the recipient was cooled to 30°C in group I, and 34°C in groups II and III. After cross-clamping the ascending aorta, the recipient's heart was excised at the atrial levels, both large vessels were cut. The porcine donor heart was transplanted using Shumway's and Lower's technique⁸.

A wireless telemetric transmitter (Data Sciences International) was implanted in a subcutaneous pouch in the right medioclavicular line between the fifth and sixth rib. Pressure probes were inserted into the ascending aorta and the apex of the left ventricle, an electrocardiogram lead was placed in the right ventricular wall.

Immunosuppressive regimen, anti-inflammatory and additive therapy. Immunosuppression was based on the previously published regimen⁵, with C1 esterase inhibitor instead of cobra venom factor for complement inhibition

(Extended Data Table 1). Induction consisted of anti-CD20 antibody (mabthera; Roche Pharma), ATG (thymoglobuline, Sanofi-Aventis), and either an anti-CD40 monoclonal antibody (mouse/rhesus chimeric IgG4 clone 2C10R4, NIH Non-human Primate Reagent Resource, Mass Biologicals; courtesy of K. Reimann; animals 1–3, 5, 7, 8, 11–14) or humanized anti-CD40L PASylated Fab (XL-Protein and Wacker-Chemie; animals 4, 6, 9, 10). Maintenance immunosuppression consisted of mycophenolate mofetil (CellCept, Roche; trough level 2–3 µg ml⁻¹), either the anti-CD40 monoclonal antibody (animals 1–3, 5, 7, 8, 11–14) or anti-CD40L PASylated Fab (animals 4, 6, 9, 10), and methylprednisolone (urbasone soluble, Sanofi-Aventis). Anti-inflammatory therapy included an IL-6-receptor antagonist (RoActemra, Roche), TNF inhibitor (enbrel, Pfizer) and an IL-1-receptor antagonist (Kineret, Swedish Orphan Biovitrum). Additive therapy consisted of acetylsalicylic acid (aspirin, Bayer Vital), unfractionated heparin (heparin-natrium-25000-ratiopharm, Ratiopharm), C1 esterase inhibitor (berinert, CSL Behring), ganciclovir (cymevene, Roche), cefuroxime (cefuroxim, Hikma) and epoetin beta (neorecormon 5000IU, Roche P).

Starting from 10 mg kg⁻¹ per day, methylprednisolone was slowly reduced by 1 mg kg⁻¹ every 10 days in group I and II; in group III, methylprednisolone was tapered down to 0.1 mg kg⁻¹ within 19 days. Also in group III, temsirolimus (torisel, Pfizer) was added to the maintenance immunosuppression, administered as daily short intravenous infusions aiming at rapamycin trough levels of 5–10 ng ml⁻¹. Group III also received continuous intravenous antihypertensive medication with enalapril (Enahexal, Hexal AG, Holzkirchen, Germany) and metoprolol tartrate (Beloc, AstraZeneca), aiming at mean arterial pressures of 80 mm Hg and a heart rate of 100 b.p.m.

Haemodynamic measurements. After induction of general anaesthesia, a central venous catheter (Arrow International) was inserted in the left jugular vein and an arterial catheter (Thermofluid Pulsio cath; Pulsion Medical Systems) in the right femoral artery. Cardiac output and stroke volume were assessed by transpulmonary thermodilution and indexed to the body surface area of the recipient using the formula $0.083 \times B^{0.639}$ where B is body weight in kg. Measurements were taken after induction of anaesthesia and 60 min after termination of CPB in steady state and recorded with PiCCOWin software (Pulsion Medical Systems). All data were processed with Excel (Microsoft) and analysed with GraphPad Prism 7.0 (GraphPad Software).

Quantification of left ventricular mass, left ventricular mass increase and fractional shortening. Transthoracic echocardiographic examinations were carried out under analgosedation at regular intervals using an HP Sonos 7500 (HP) and a Siemens Acuson X300 (Siemens); midpapillary short axis views were recorded. Left ventricular end diastolic diameter (LVEDD) and left ventricular end systolic diameter (LVESD), interventricular septum thickness at end diastole (IVSd) and posterior wall thickness at end diastole (PWd) were measured; the mean of three measurements was used for further calculations and visualization (Excel and PowerPoint, Microsoft).

Left ventricular mass was calculated using equation (1), relative left ventricular mass increase and left ventricular (LV) fractional shortening (FS) was calculated using equations (2) and (3) according to previously published methods^{14,15}.

$$\text{LV mass (g)} = 0.8(1.04((\text{LVEDD} + \text{IVSd} + \text{PWd})^3 - \text{LVEDD}^3)) + 0.6 \quad (1)$$

$$\text{LV mass increase (\%)} = ((\text{LV mass}_{\text{end}}/\text{LV mass}_{\text{start}}) - 1) \times 100 \quad (2)$$

$$\text{FS (\%)} = ((\text{LVEDD} - \text{LVESD})/\text{LVEDD}) \times 100 \quad (3)$$

Necropsy and histology. Necropsies and histology were performed at the Institute of Veterinary Pathology and the Institute of Pathology. Specimens were fixed in formalin, embedded in paraffin and plastic, sectioned and stained with haematoxylin and eosin.

Histochemical analysis. Cryosections (8 µm) were generated using standard histological techniques. Cardiomyocyte size was quantified as the cross-sectional area. In brief, 8-µm thick cardiac sections of the left ventricle were stained with Alexa Fluor 647-conjugated wheat germ agglutinin (Life Technologies) and the nuclear dye 4',6-diamidino-2-phenylindole (DAPI, Life Technologies). Images were acquired with a 63× objective using a Leica TCS SP8 confocal microscope; SMASH software (MATLAB, <https://de.mathworks.com/products/matlab.html>) was used to determine the average cross-sectional area of cardiomyocytes in one section (200–300 cells per section and 5–8 sections per heart).

Immunofluorescence staining. Myocardial tissue biopsies were embedded in Tissue-Tek (Sakura Finetek) and stored frozen at –80°C. For immunofluorescence staining, 5-µm cryosections were cut, air-dried for 30 to 60 min and stored at –20°C until further analysis. The cryosections were fixed with ice-cold acetone, hydrated and stained using either one-step direct or two-step indirect immunofluorescence techniques. The following antibodies were used: rabbit anti-human C3b/c (DAKO), rabbit anti-human C4b/c-FITC (DAKO), goat anti-pig IgM (AbD

Serotec), goat anti-human IgG–FITC (Sigma–Aldrich) and rabbit anti-human fibrinogen–FITC (DAKO). Secondary antibodies were donkey anti-goat IgG–Alexa Fluor 488 (Thermo Fisher Scientific), sheep anti-rabbit Cy3 (Sigma–Aldrich). Nuclear staining was performed using DAPI (Boehringer, Roche Diagnostics). The slides were analysed using a fluorescence microscope (DM14000B; Leica). Five to ten immunofluorescence pictures per marker were acquired randomly and the fluorescence intensity was quantified using ImageJ software, version 1.50i (<https://imagej.nih.gov/ij/>) on unmanipulated TIFF images. All pictures were taken under the same conditions to allow correct quantification and comparison of fluorescence intensities.

Assessment of non-galactose- α 1,3-galactose antibody levels. Plasma levels of non-galactose- α 1,3-galactose baboon IgM and IgG antibodies were measured by flow cytometry following the consensus protocol published previously¹⁷. In brief, GTKO/hCD46/hTM porcine aortic endothelial cells were collected and suspended at 2×10^6 cells per ml in staining buffer (PBS containing 1% BSA). Plasma samples were heat-inactivated at 56 °C for 30 min and diluted 1:20 in staining buffer. Porcine aortic endothelial cells were incubated with diluted baboon plasma for 45 min at 4 °C. Cells were then washed with cold staining buffer and incubated with goat anti-human IgM–RPE (Southern Biotech) or goat anti-human IgG–FITC (Thermo Fisher) for 30 min at 4 °C. After rewashing with cold staining buffer, cells were resuspended in PBS, fluorescence was acquired on FACS LSRII (BD Biosciences) and data were analysed using FlowJo analysis software for detection of mean fluorescence intensity (MFI) in the FITC channel or in the RPE channel. Data were then plotted using Prism 7 (GraphPad Software).

Western blot analysis. For protein extraction, heart samples were homogenized in Laemmli sample buffer and the protein content was estimated using the bicinchoninic acid (BCA, Merck) protein assay. Then, 20 μ g total protein was separated by 10% SDS–PAGE and transferred to PDVF membranes (Millipore) by electroblotting. Membranes were washed in Tris-buffered saline solution with 0.1% Tween-20 (Merck) (TBS–T) and blocked in 5% w/v fat-free milk powder (Roth) for 1 h at room temperature. Membranes were then washed again in TBS–T and incubated in 5% w/v BSA (Roth) of the appropriate primary antibody overnight at 4 °C. The following antibodies were used: rabbit anti-human p-mTOR (5536; Cell Signaling), rabbit anti-human mTOR (2983; Cell Signaling) and rabbit anti-human GAPDH (2118; Cell Signaling). After washing, membranes were incubated

in 5% w/v fat-free milk powder with a horseradish peroxidase-labelled secondary antibody (goat anti-rabbit IgG; 7074; Cell Signaling) for 1 h at room temperature. Bound antibodies were detected using an enhanced chemiluminescence detection reagent (ECL Advance Western Blotting Detection Kit, GE Healthcare) and appropriate X-ray films (GE Healthcare). After detection, membranes were stripped (2% SDS, 62.5 mM Tris–HCl, pH 6.7, 100 mM β -mercaptoethanol) for 30 min at 70 °C and incubated with the appropriate second antibody.

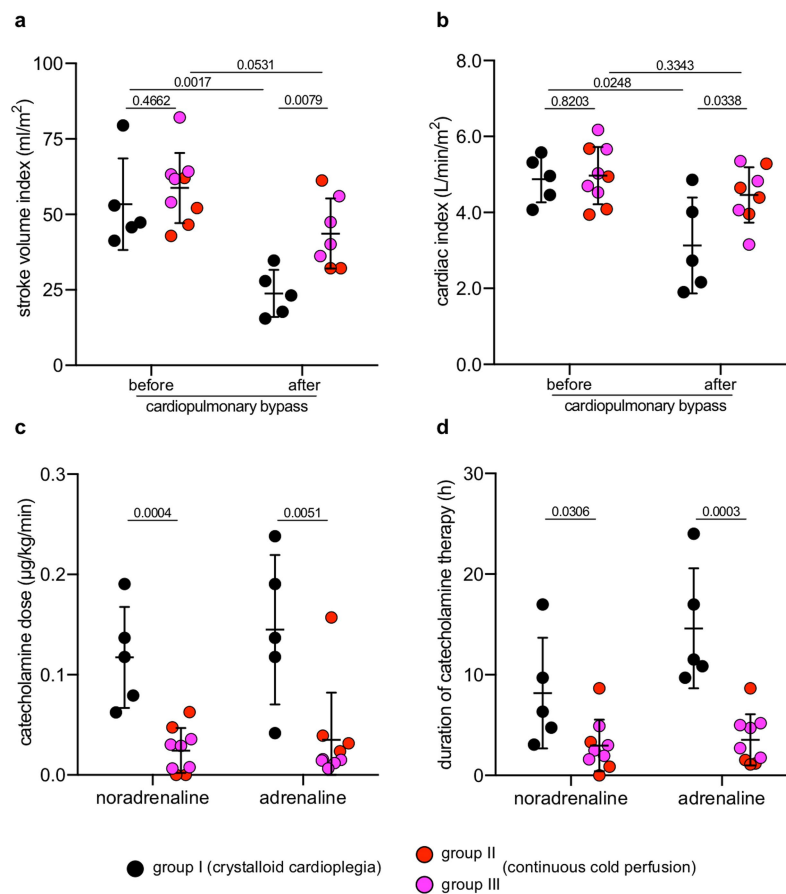
Immunohistochemical staining. Myocardial tissue was fixed with 4% formalin overnight, paraffin-embedded and 3- μ m sections were cut and dried. Heat-induced antigen retrieval was performed in Target Retrieval solution (S1699, DAKO) in a boiling water bath for 20 min for hCD46 and in citrate buffer, pH 6.0, in a steamer for 45 min for hTM, respectively. Immunohistochemistry was performed using the following primary antibodies: mouse anti-human CD46 monoclonal antibody (HM2103, Hycult Biotech) and mouse anti-human thrombomodulin monoclonal antibody (sc-13164, Santa Cruz). The secondary antibody was a biotinylated AffiniPure goat anti-mouse IgG (115-065-146, Jackson ImmunoResearch). Immunoreactivity was visualized using 3,3'-diaminobenzidine tetrahydrochloride dihydrate (DAB) (brown colour). Nuclear counterstaining was done with haemalum (blue colour).

Statistical analysis. For survival data, Kaplan–Meier curves were plotted and the Mantel–Cox log-rank test was used to determine significant differences between groups. For haemodynamic data, statistical significance was determined using unpaired and paired two-sided Student's *t*-tests as indicated; data presented as single measurements with bars as group means \pm s.d. For histochemical analysis, a one-way ANOVA with Holm–Sidak's multiple comparisons was used to determine statistical significance; data are mean \pm s.d.; $P < 0.05$ was considered significant. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

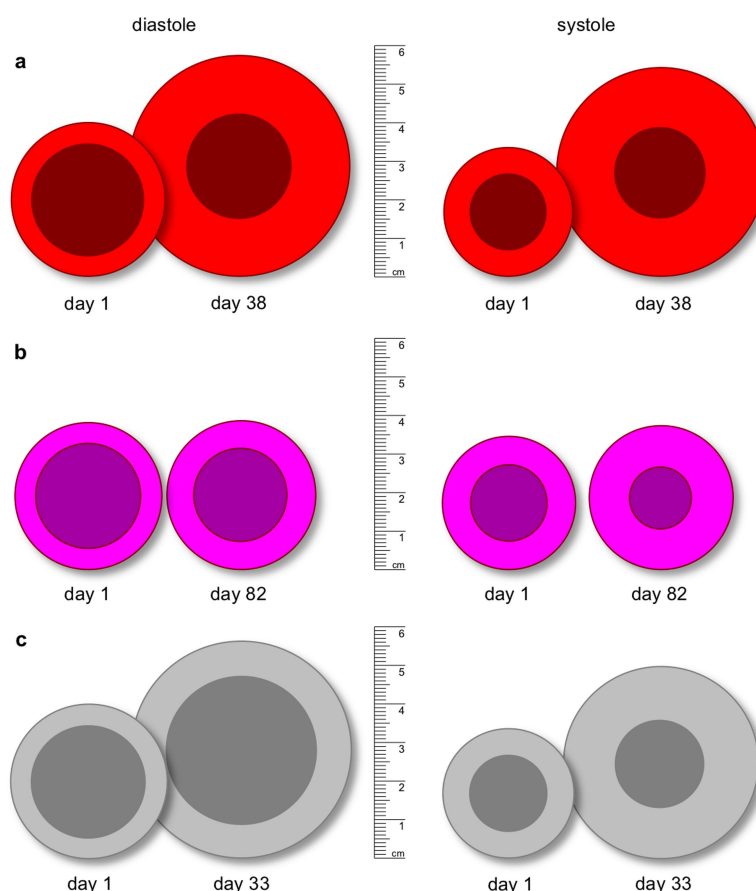
Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.



Extended Data Fig. 1 | Haemodynamic data, measured by transpulmonary thermodilution and post-operative catecholamine support. Measurements were taken after induction of anaesthesia (before) and 60 min after termination of CPB (after). Donor hearts of group I (black) received crystalloid cardioplegia, donor hearts of groups II (red) and III (magenta) were preserved with continuous cold hyperoncotic perfusion; data are presented as scatter plots with mean \pm s.d. with individuals shown as dots; $n = 14$ animals, two-sided paired and

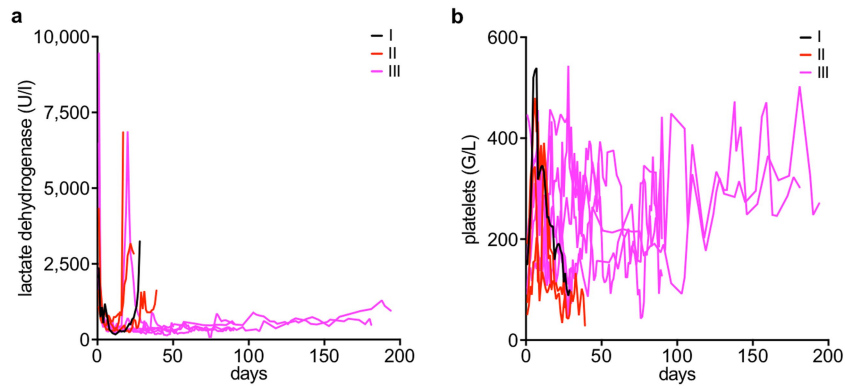
unpaired *t*-tests, *P* values as indicated. **a**, Stroke volume index. **b**, Cardiac index before and after CPB. Both parameters decreased in group I and were lower in group I after CPB than in group II and III. **c**, Dosages of catecholamines 60 min after termination of CPB. **d**, Durations of post-operative vasopressive and inotropic support. Animals in group I required more noradrenaline and adrenaline than those in group II and III. Animals in group I required inotropic support with adrenaline for a longer time.



Extended Data Fig. 2 | Graphics of left ventricular sizes during diastole and systole that were derived from transthoracic echocardiography.

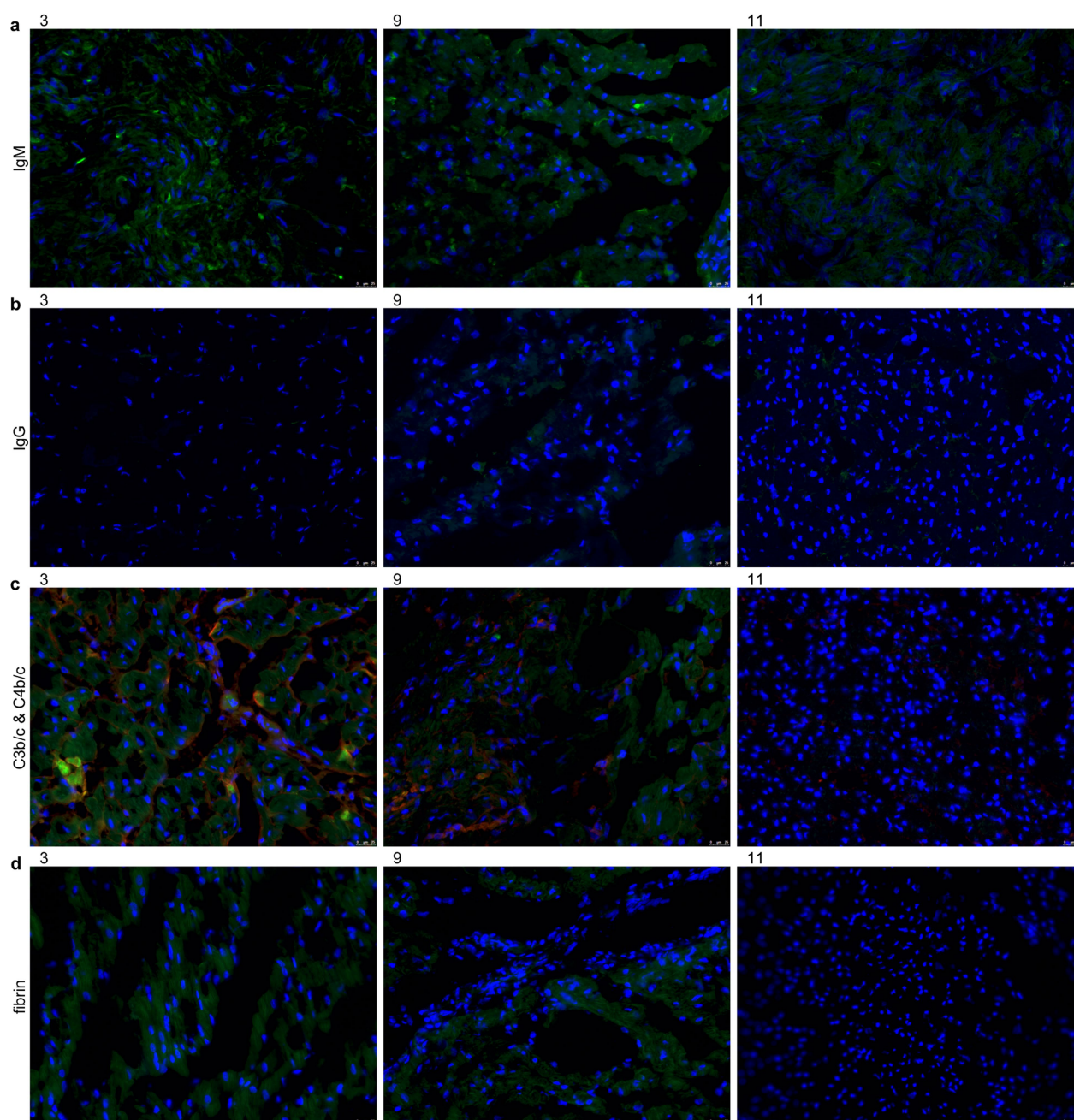
Left, diastole; right, systole. **a**, Animal 9 (group II, survival 40 days): left ventricular mass had increased by 303% on day 38, left ventricular function was severely impaired because of myocardial hypertrophy and decreased left ventricular filling volume. Left ventricular fractional shortening measurements were 32% and 14% on day 1 and 38. **b**, Animal 11 (group III, survival 90 days): in contrast to animal 9, left ventricular

mass had increased by only 22% on day 82, left ventricular function was preserved. Left ventricular fractional shortening measurements were 27% and 34% on day 1 and 82. **c**, Pig 5157 (control, donor sibling of the pig whose heart was transplanted in animal 9): left ventricular mass had increased by 187% on day 33, left ventricular function was preserved. Left ventricular fractional shortening measurements were 32% and 41% on day 1 and 33. Compared to 9 (**a**), the left ventricle had grown less in size and showed no hypertrophy.



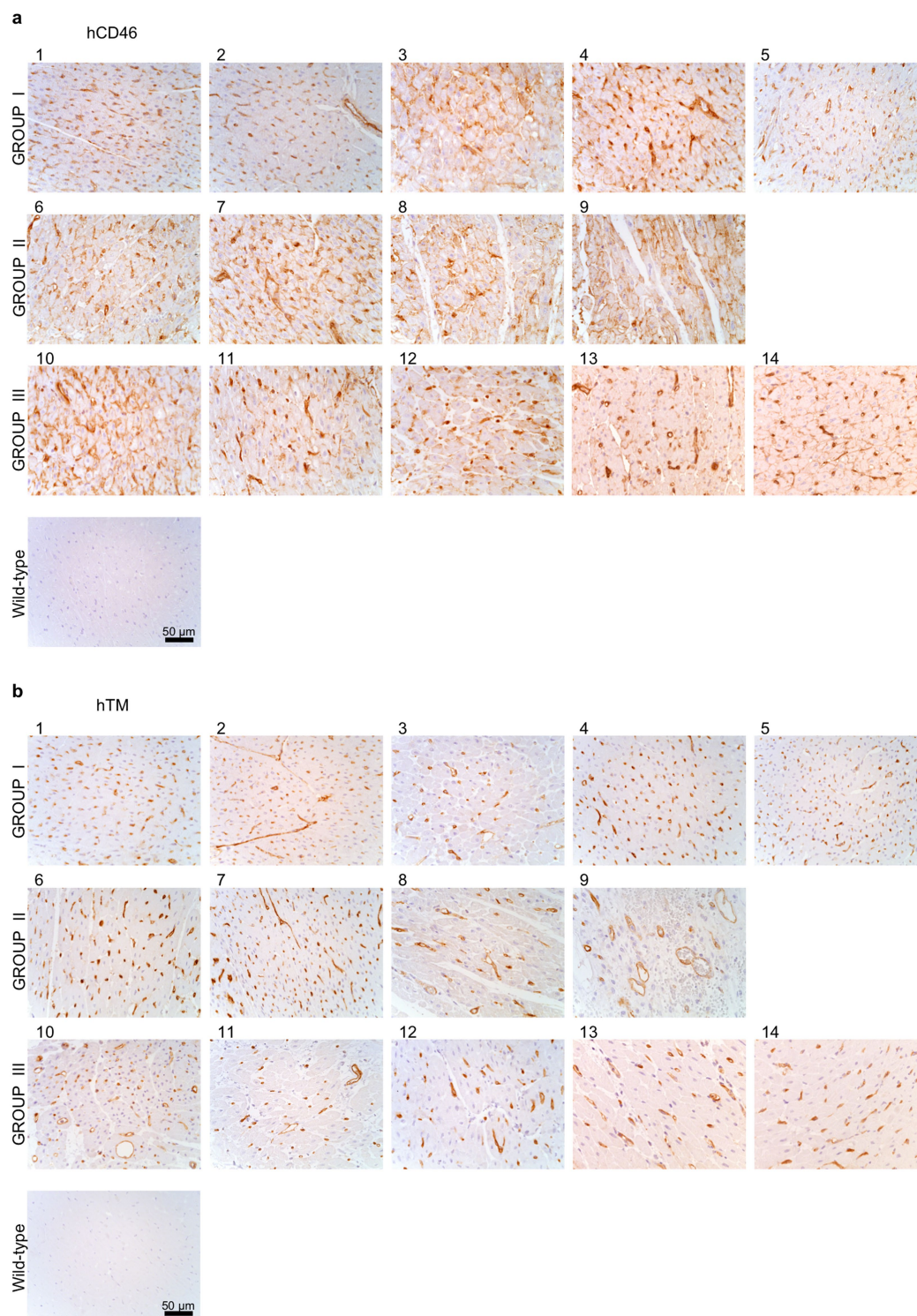
Extended Data Fig. 3 | Additional laboratory parameters. a, b, Serum concentrations of lactate dehydrogenase (a) and platelet counts (b) in animals of groups I (black), II (red) and III (magenta). At the end of

experiments in groups I and II, platelet counts decreased whereas LDH increased. Group III animals did not show these alterations.



Extended Data Fig. 4 | Immunofluorescence staining of myocardial tissue. **a–d**, Immunofluorescence staining of myocardial sections from group I (3; left row), group II (9; middle row) and group III (11, right row) for IgM (**a**), IgG (**b**), C3b/c (c; red), C4b/c (c; green) and fibrin (**d**); nuclei

were stained with DAPI (blue). Scale bars, 25 μm . $n = 1$, group I; $n = 3$, group II; $n = 5$, group III; one representative biological sample per group is shown.



Extended Data Fig. 5 | Immunohistochemistry of post-mortem myocardial specimens. a, b, Expression of human membrane cofactor protein (hCD46) (**a**) and human thrombomodulin (hTM) (**b**) was

consistent in all donor organs (1–14). Scale bars, 50 μ m. $n = 14$ GTKO/hCD46/hTM pigs; $n = 1$ wild-type pig (control). Biological samples from all animals are shown.

Extended Data Table 1 | Immunosuppressive regimen, anti-inflammatory and additive therapy with corresponding doses and timing intervals

| Agent | Dose | Timing |
|--|--|------------------------------------|
| Induction | | |
| anti-CD20 Ab | 19 mg kg ⁻¹ , i.v. short infusion | days -7, 0, 7 and 14 |
| ATG | 5 mg kg ⁻¹ , continuously i.v. | days -2 and -1 |
| anti-CD40 mAb or anti-CD40L PASylated Fab* | 50 mg kg ⁻¹ or 20 mg kg ⁻¹ ; i.v. short infusion | days -1 and 0 |
| Maintenance | | |
| MMF | 40 mg kg ⁻¹ , continuously i.v. | daily, started on day -2 |
| anti-CD40 mAb or anti-CD40L PASylated Fab* | 50 mg kg ⁻¹ or 20 mg kg ⁻¹ i.v. short infusion | days 3, 7, 10, 14, 19, then weekly |
| methylprednisolone | 10 mg kg ⁻¹ , bolus i.v. | daily, tapered down |
| Anti-inflammatory therapy | | |
| IL6-receptor antagonist | 8 mg kg ⁻¹ , short infusion i.v. | monthly |
| TNF α inhibitor | 0.7 mg kg ⁻¹ , bolus s.c. | weekly |
| IL1-receptor antagonist | 1.3 mg kg ⁻¹ , bolus s.c. or i.v. | daily |
| Additive therapy | | |
| acetylsalicylic acid | 2 mg kg ⁻¹ , bolus i.v. | daily |
| unfractionated heparin | 20–40 U kg ⁻¹ h ⁻¹ , continuously i.v. | daily, started on day 5 |
| C1 esterase inhibitor | 17.5 U kg ⁻¹ , i.v. short infusion | days 0, 1, 7 and 14 |
| ganciclovir | 5 mg kg ⁻¹ , continuously i.v. | daily |
| cefuroxim | 50 mg kg ⁻¹ , continuously i.v. | daily, prophylaxis from day 0 to 5 |
| epoetin beta | 2,000 U, bolus s.c. or i.v. | days -7, 0 and if necessary |

Immunosuppression was based on the previously published regimen⁵, with C1 esterase inhibitor instead of cobra venom factor for complement inhibition. Starting from 10 mg kg⁻¹ per day, methylprednisolone was slowly reduced by 1 mg kg⁻¹ every 10 days in group I and II; in group III, methylprednisolone was reduced to 0.1 mg kg⁻¹ within 19 days. In group III, temsirolimus was added to the maintenance immunosuppression, administered as daily infusions (rapamycin trough levels: 5–10 ng ml⁻¹). Group III animals also received continuous antihypertensive medication (enalapril and metoprolol tartrate). Ab, antibody; mAb, monoclonal antibody; ATG, anti-thymocyte globulin; CMV, cytomegalovirus; IgG4, immunoglobulin G4; IL, interleukin; i.v., intravenous; MMF, mycophenolate mofetil; PASylated, conjugated with a long structurally disordered Pro-Ala-Ser amino acid chain; s.c., subcutaneous; TNF α , tumour necrosis factor α .

*Anti-CD40 monoclonal antibody: animals 1–3, 5, 7, 8, 11–14; anti-CD40L PASylated Fab: animals 4, 6, 9, 10.

The translation of non-canonical open reading frames controls mucosal immunity

Ruaidhrí Jackson¹, Lina Kroehling¹, Alexandra Khitun², Will Bailis¹, Abigail Jarret¹, Autumn G. York¹, Omair M. Khan¹, J. Richard Brewer¹, Mathias H. Skadow¹, Coco Duizer¹, Christian C. D. Harman¹, Lelina Chang¹, Piotr Bielecki¹, Angel G. Solis¹, Holly R. Steach¹, Sarah Slavoff^{2,3,4} & Richard A. Flavell^{1,5*}

The annotation of the mammalian protein-coding genome is incomplete. Arbitrary size restriction of open reading frames (ORFs) and the absolute requirement for a methionine codon as the sole initiator of translation have constrained the identification of potentially important transcripts with non-canonical protein-coding potential^{1,2}. Here, using unbiased transcriptomic approaches in macrophages that respond to bacterial infection, we show that ribosomes associate with a large number of RNAs that were previously annotated as ‘non-protein coding’. Although the idea that such non-canonical ORFs can encode functional proteins is controversial^{3,4}, we identify a range of short and non-ATG-initiated ORFs that can generate stable and spatially distinct proteins. Notably, we show that the translation of a new ORF ‘hidden’ within the long non-coding RNA *Aw112010* is essential for the orchestration

of mucosal immunity during both bacterial infection and colitis. This work expands our interpretation of the protein-coding genome and demonstrates that proteinaceous products generated from non-canonical ORFs are crucial for the immune response in vivo. We therefore propose that the misannotation of non-canonical ORF-containing genes as non-coding RNAs may obscure the essential role of a multitude of previously undiscovered protein-coding genes in immunity and disease.

Ribosome association with mRNA is essential for protein synthesis⁵. Here we investigated the genome-wide association of mRNA with ribosomes in macrophages after bacterial infection by generating RiboTag-LysM-Cre (RiboTag^{LysM}) mice^{6,7}. Bone marrow-derived macrophages (BMDMs) from wild-type and RiboTag^{LysM} mice were generated and stimulated with lipopolysaccharide (LPS) for 6 or 24 h.

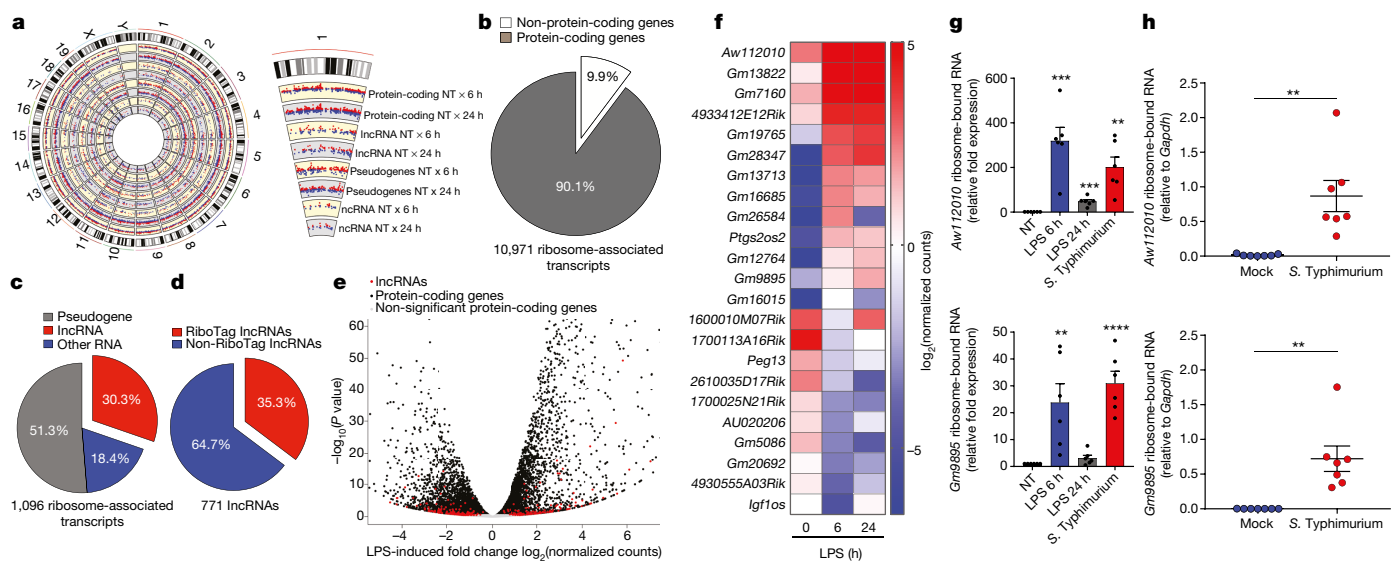


Fig. 1 | Bacterial infection drives widespread ribosomal association with non-coding RNAs. **a–f**, BMDMs from RiboTag^{LysM} mice were non-treated (NT) or stimulated with LPS (1 ng ml⁻¹). RNA was subjected to RNA-seq. Data are presented as a combination of two independent biological replicates. **a**, Circos plot shows differentially expressed ribosome-associated transcripts after 6 and 24 h LPS stimulation. Red denotes upregulation; blue denotes downregulation. Each track from the periphery to the core represents: chromosomes location; 12,820 known protein-coding transcripts; 1,176 lncRNAs; 1,107 pseudogenes; and 413 other non-coding RNAs. **b**, Pie chart of the percentage breakdown of protein-coding genes annotated from RiboTag RNA-seq (fragments per kilobase of transcript per million mapped reads (FPKM) ≥ 1). **c**, The non-protein coding genes in **b** are further classified. **d**, Stratification of detectable BMDM lncRNAs based on ribosome association. Ribosome-associated lncRNAs with an

FPKM of ≥ 1 in RiboTag RNA-seq are represented in the red exploded section. Blue section depicts lncRNAs not found in RiboTag RNA-seq, but with an FPKM of ≥ 0.01 in conventional RNA-seq. **e**, **f**, Volcano plot (**e**) and heat map analysis (**f**) of lncRNAs associated with ribosomes after LPS stimulation in BMDMs. **g**, qPCR analysis of ribosome-associated transcripts of non-treated BMDMs or those stimulated with LPS (10 ng ml⁻¹) or infected with *S. Typhimurium* at a multiplicity of infection (MOI) of 1 for 6 h. Data are presented as six biological replicates, and fold expression was calculated from each individual non-treated sample. **h**, RiboTag^{LysM} mice were gavaged with 2×10^8 CFUs of *S. Typhimurium*. After 24 h, colonic tissue was extracted and lysed. Macrophage ribosome-associated RNA was isolated and qPCR analysis was conducted. Data are presented as seven biological replicates. Data are mean and s.e.m. ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$, unpaired two-tailed t -test.

¹Department of Immunobiology, Yale University School of Medicine, New Haven, CT, USA. ²Department of Chemistry, Yale University, New Haven, CT, USA. ³Chemical Biology Institute, Yale University, West Haven, CT, USA. ⁴Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. ⁵Howard Hughes Medical Institute, Yale University, New Haven, CT, USA. *e-mail: richard.flavell@yale.edu

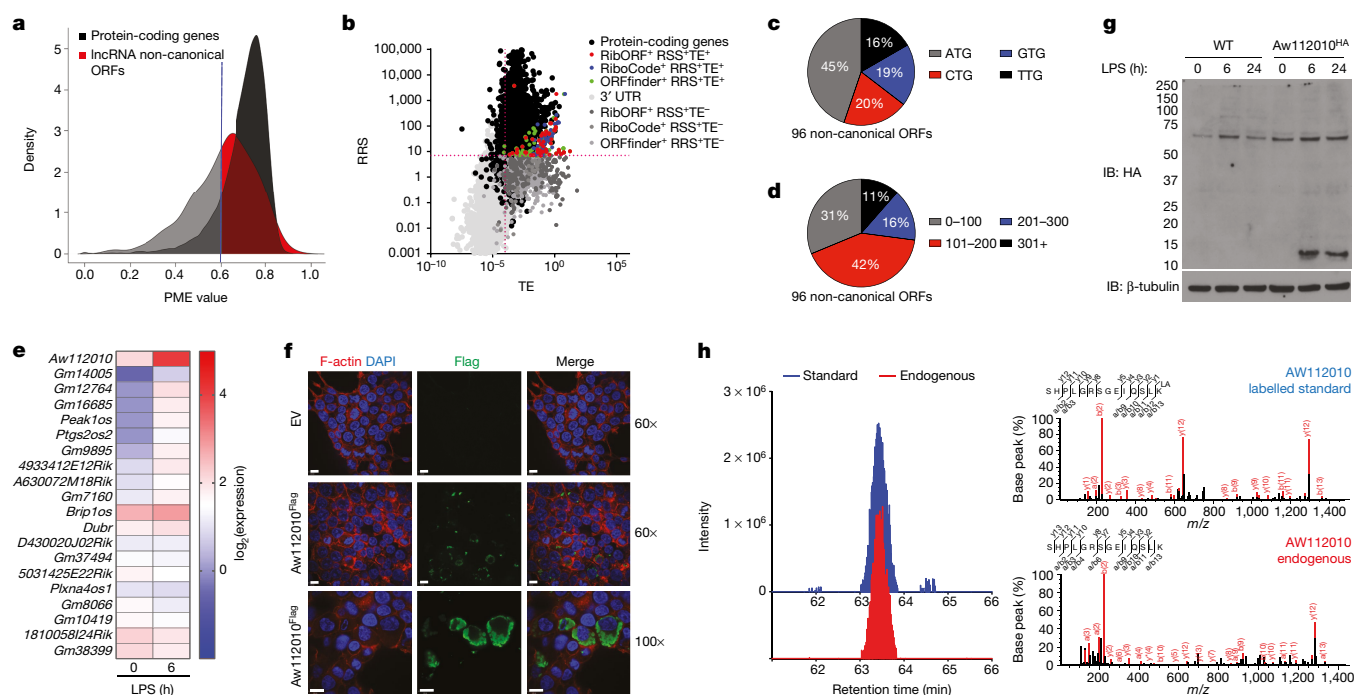


Fig. 2 | LPS triggers genome wide differential translation of non-canonical ORFs in lncRNAs. **a–e**, Wild-type BMDMs were non-treated or stimulated with LPS (10 ng ml⁻¹) for 6 h and ribosome profiling was conducted. Data are representative of two biological replicates. **a**, PME values for protein-coding genes and lncRNAs. A PME cut-off value of $\geq 0.6\%$ represents transcripts considered to be protein coding. **b**, Translation efficiency (TE) and RRS analysis was conducted on transcripts identified by RiboProfiling. Purple broken lines represent the 95th percentile of the 3' UTRs of known protein-coding genes and discriminates coding and non-coding transcripts. **c**, **d**, Categorization of lncRNAs with coding RRS and translation efficiency values. **e**, Heat map of the top differentially regulated LPS-stimulated lncRNA ORFs. **f**, HEK293 cells transfected with empty vector (EV) or

Aw112010^{Flag} ORF. Cells were stained with DAPI, phalloidin and anti-Flag. Scale bars, 9 μm; original magnifications, ×60 and ×100. **g**, Wild-type (WT) and Aw112010^{HA} BMDMs were non-treated or stimulated with LPS (10 ng ml⁻¹), protein lysates were generated and western blot analysis was conducted for haemagglutinin and β-tubulin. Data are representative of three biological replicates. **h**, Aw112010^{HA} BMDMs were generated, stimulated with LPS for 6 h and subjected to haemagglutinin immunoprecipitation. Purified lysates were subject to mass spectrometry analysis. Precursor ion peaks in the MS1 extracted ion chromatogram corresponding to a spiked in synthetic isotopically labelled peptide standard (top) and co-elution of a peak consistent with the endogenous Aw112010 peptide (bottom) in the same sample. Identified fragment ions (b and y ions, red) are indicated above and below the peptide sequence. Data are representative of two biological replicates.

Immunoprecipitation using haemagglutinin-conjugated magnetic beads and subsequent RNA isolation yielded high-quality RNA from BMDMs from RiboTag^{LysM} mice and no detectable ribosome-associated RNA in wild-type cells (Extended Data Fig. 1a, b). RiboTag RNA sequencing (RNA-seq) revealed widespread differential ribosome assembly on both protein-coding RNAs and, unexpectedly, on transcripts mapping to 'non-coding' RNAs^{8–10} (Fig. 1a). In fact, almost 10% of all ribosome-associated RNAs were annotated as 'non-coding', with pseudogenes and long non-coding RNAs (lncRNAs) representing the most abundant classes (Fig. 1b, c). lncRNAs have been described to have major functions in diverse biological systems, including immune cell development and function^{11,12}. According to canonical ORF designations, lncRNAs are unable to code for protein^{11,12}; however, we found that more than 35% of highly expressed macrophage lncRNAs interact with ribosomes during bacterial infection (Fig. 1d). To identify potentially important protein-coding lncRNAs in the innate immune response, we identified genes that were markedly altered after LPS stimulation (Fig. 1e, f). We confirmed ribosome association with candidate lncRNAs in RiboTag^{LysM} BMDMs stimulated with LPS or infected with *Salmonella enterica* serovar Typhimurium (*S. Typhimurium*) using quantitative PCR (qPCR). Both LPS stimulation and *S. Typhimurium* infection induced differential ribosome association and transcription of these genes (Fig. 1g and Extended Data Fig. 1c). Finally, we infected RiboTag^{LysM} mice with *S. Typhimurium* and isolated ribosome-associated RNA from colonic macrophages 24 h after bacterial gavage. qPCR analysis revealed significant increases in ribosome-association for the lncRNAs *Aw112010* and *Gm9895* after infection (Fig. 1h).

Although RiboTag RNA-seq reveals widespread association of lncRNAs with ribosomes during bacterial infection, ribosome association per se does not necessarily indicate whether a given RNA is being actively translated¹³. Ribosome-profiling techniques have emerged as powerful tools to address such caveats¹⁴. Using steady-state and LPS-stimulated wild-type BMDMs, we generated genome-wide ribosome profiles in tandem with conventional poly(A)⁺ RNA-seq. This allowed the successful discrimination of the known protein-coding genes and non-coding RNAs (Extended Data Fig. 2a, b). Notably, however, we also identified a plethora of lncRNAs with distinct ribosome profiles that were similar to that of known protein-coding genes (Extended Data Fig. 2c). We next sought to identify actively translated ORFs in an unbiased manner and conducted RibORF analysis¹⁵. RibORF correctly identified transcripts undergoing active translation (Extended Data Fig. 2d). During the classical annotation of the genome, protein-coding genes were accurately identified by the presence of an ATG methionine start codon and an ORF of greater than 300 nucleotides^{16,17}. However, proteinaceous products with as few as 11 amino acids¹⁸ and single nucleotide promiscuity in near cognate ATG codons that facilitate translation initiation¹⁹ have been reported. Using a custom ORFinder search to identify all ORFs that were more than 30 nucleotides long using the start codons ATG, CTG, TTG or GTG²⁰, we generated a library of all potential non-canonical ORFs within BMDM-expressed lncRNAs. RibORF analysis identified 224 non-canonical ORFs with the same translation hallmarks (percentage of maximum entropy (PME) ≥ 0.6 ; ref. ²¹) as protein-coding genes, and predicted that they undergo active translation (Fig. 2a). In addition, we wished to identify ORFs using a method that does not require the ORF to be initially

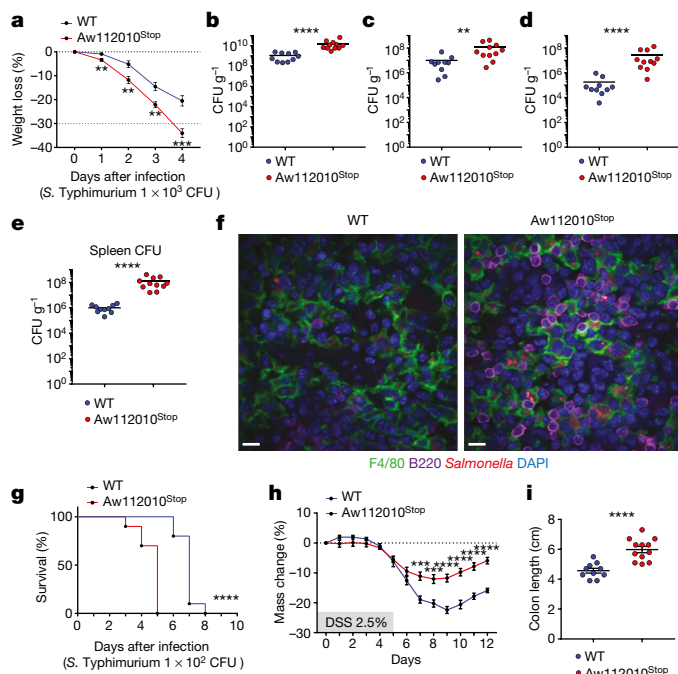


Fig. 3 | Translation of the non-canonical Aw112010 encoded ORF is essential for mucosal immunity. **a–e**, Wild-type ($n = 10$) and Aw112010^{Stop} ($n = 11$) mice were administered streptomycin (20 mg) by oral gavage 24 h before *S. Typhimurium* infection (1×10^3 CFUs). **a**, Weight loss was measured after infection. **b**, Enumeration of *S. Typhimurium* CFUs present in the faeces of wild-type and Aw112010^{Stop} mice 24 h after infection. **c**, Enumeration of *S. Typhimurium* CFUs in the caecum of wild-type and Aw112010^{Stop} mice 96 h after infection. **d**, **e**, Enumeration of *S. Typhimurium* CFUs in the liver (**d**) and spleen (**e**) of wild-type and Aw112010^{Stop} mice 96 h after infection. **f**, Confocal immunostaining of macrophages (F4/80, green), B cells (B220, purple) and *Salmonella* (anti-*Salmonella*, red) in the spleens of wild-type and Aw112010^{Stop} mice infected with 1×10^2 CFUs of *S. Typhimurium* 72 h after gavage. Images are representative of three independent biological replicates. Scale bars, 9 μm; original magnification, $\times 60$. **g**, Survival curve analysis of wild-type ($n = 10$) and Aw112010^{Stop} ($n = 10$) mice infected with 1×10^2 CFUs via oral gavage. **h**, **i**, Wild-type and Aw112010^{Stop} cohoused littermate mice were administered 2.5% DSS in their drinking water for 5 days. **h**, Weight loss from wild-type ($n = 11$) and Aw112010^{Stop} ($n = 12$) mice was measured over 12 days. **i**, Colon length was measured from wild-type ($n = 10$) and Aw112010^{Stop} ($n = 12$) mice. Horizontal bars in **b–e** represent the mean bacterial count. Error bars in **h** and **i** denote s.e.m. of replicates. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$, nonparametric Mann–Whitney test (**b–e**), log-rank test (**g**), and unpaired two tailed t -test (**a**, **h**, **i**).

defined with ORFfinder. RiboCode analysis²² identified 85 non-canonical ORFs within lncRNAs de novo (Extended Data Fig. 3a). As previously reported²², RibORF and RiboCode analyses identify both similar and unique ORFs (Extended Data Fig. 3b). To investigate the translational veracity of these ORFs further, we conducted another analytical strategy previously used to challenge the assertion that lncRNAs can encode translated proteins. Using translational efficiency⁴ and ribosome release score (RRS) analysis to both the 3' untranslated regions (UTRs) and the known protein-coding ORF within these genes, we generated a frame of reference for macrophage-translated and non-translated transcripts. Using the ninety-fifth percentiles of the 3' UTR translational efficiency and RRS, we can predict the ability of a given ORF to encode a translated protein (Fig. 2b). By selecting non-canonical ORFs identified by RibORF and RiboCode analysis, we identified 96 lncRNAs that share the RRS and translational efficiency values with those of the known protein-coding genome. Nearly half of these ORFs use the ATG start codon, whereas the remaining ORFs shared a relatively even distribution of start codon usage between TTG,

GTG and CTG (Fig. 2c). Most are small ORFs that are under 300 nucleotides in length (Fig. 2d). To identify putative ORFs that can encode functionally important proteins in the immune response, we performed differential gene expression analysis on these non-canonical ORFs in LPS-stimulated macrophages from the ribosome-profiling sequencing dataset (Fig. 2e and Extended Data Fig. 3c). Aw112010 was significantly upregulated by LPS, and because it was identified as protein coding by RRS, translational efficiency, RibORF and RiboCode analysis, we choose to investigate this non-canonical ORF further. Overexpression of the Aw112010 ORF in HEK293 cells demonstrated robust protein expression and distinct subcellular localization (Fig. 2f), whereas overexpression of other identified ORFs, *Gm9895* and *Gm7160*, showed strong vesicular localization and predominantly cytoplasmic staining (Extended Data Fig. 4). As such overexpression systems are artificial and cannot definitively establish real translation of an ORF in its natural context, we wished to determine whether translation of Aw112010 occurs endogenously in response to LPS. As no antibodies exist for this uncharacterized ORF, we generated an epitope-tagged Aw112010 mouse. RRS analysis for Aw112010 identified the native translation termination codon, and using a guide RNA that targeted this locus and a donor single-stranded DNA (ssDNA) template, we successfully introduced a C-terminal haemagglutinin-epitope tag into the Aw112010 gene in mice using CRISPR–Cas9 technology (Extended Data Fig. 5a–c). Generation of BMDMs from wild-type and Aw112010^{HA} mice revealed that the Aw112010 'lncRNA' does in fact generate a stable protein in response to LPS even 24 h after stimulation (Fig. 2g). We next sought to characterize in more detail the haemagglutinin-tagged protein observed in the Aw112010^{HA} macrophages. To enrich for the endogenous protein, we stimulated Aw112010^{HA} BMDMs for 6 h with LPS, and cell lysates were generated and subjected to anti-haemagglutinin immunoprecipitation (Extended Data Fig. 5d). Aw112010^{HA}-purified fractions were subjected to mass spectrometry²³. Endogenous Lys-C protease-digested peptides mapped uniquely to the predicted Aw112010 ORF-encoded protein with more than 50% total protein coverage (Extended Data Fig. 6a, b). Furthermore, we validated one of the peptide assignments with an isotopically labelled standard (labelled with ¹⁵N₂¹³C₆-lysine at the C-terminal residue), which showed co-elution and co-fragmentation with the unlabelled endogenous peptide from the stimulated Aw112010^{HA} macrophage proteome (Fig. 2h and Extended Data Fig. 6c). Together, these data indicate that non-canonical ORFs can generate abundant and stable proteins that exhibit discrete subcellular localizations. Furthermore, we demonstrate that Aw112010 is a bona fide non-canonical ORF protein-coding gene that is translated during the innate immune response to bacterial infection.

To investigate whether the translation of the non-canonical ORF within Aw112010 was physiologically important in the immune response, we generated Aw112010^{Stop} knock-in mice. We used CRISPR–Cas9 technology to insert a small frameshifting stop cassette sequence into an area of high ribosomal occupancy to abrogate its protein-coding potential (Extended Data Fig. 7a–d). Because bacterial infection can induce robust translation of Aw112010, we infected wild-type and littermate Aw112010^{Stop} mice with 1×10^3 colony forming units (CFUs) of *S. Typhimurium* via oral gavage. Aw112010^{Stop} mice displayed accelerated weight loss compared to wild-type littermates (Fig. 3a). Indeed, disruption of the Aw112010 ORF resulted in increased faecal CFUs 24 h after infection (Fig. 3b). Similarly, mice euthanized 4 days after infection showed increased bacterial load in the caecum of Aw112010^{Stop} mice (Fig. 3c). Furthermore, Aw112010^{Stop} mice presented with higher bacterial burden and dissemination to the liver and spleen than wild-type littermates (Fig. 3d, e). In addition, when Aw112010^{Stop} mice were infected with 1×10^2 CFUs of *S. Typhimurium*, they displayed accelerated bacterial dissemination and became moribund with bacterial infection significantly quicker than wild-type mice (Fig. 3f, g). To investigate whether Aw112010 also contributed to mucosal auto-inflammatory disorders such as models of inflammatory bowel disease, we administered 2.5% dextran sulfate sodium (DSS) to the drinking water of wild-type and Aw112010^{Stop}

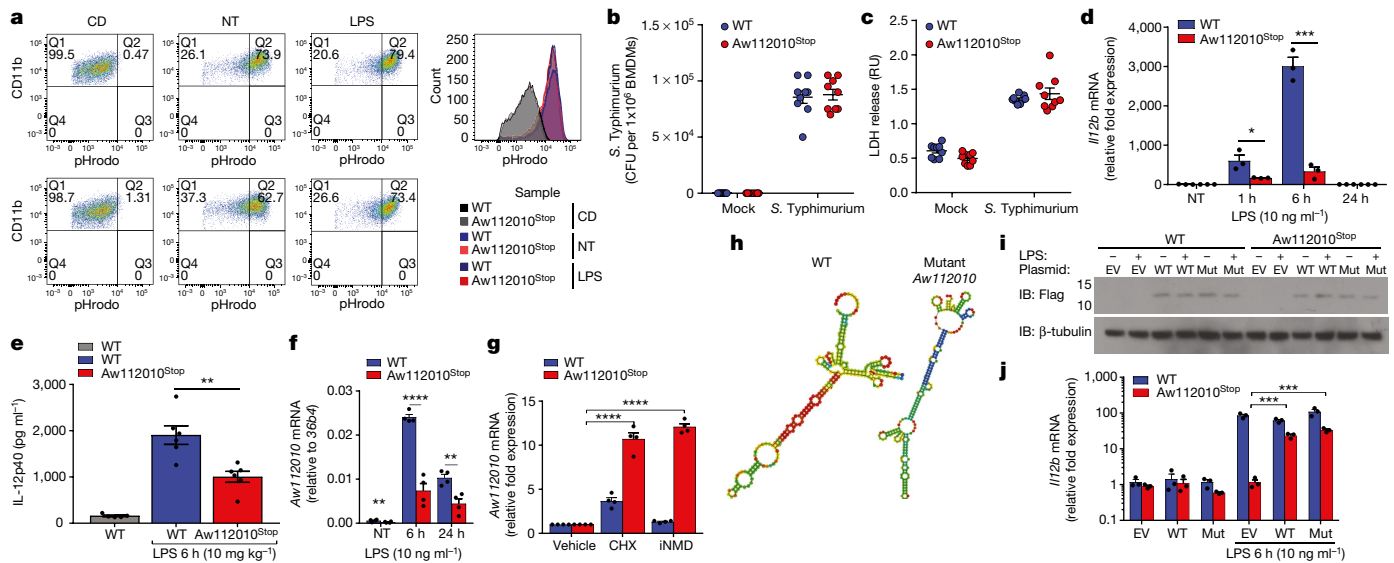


Fig. 4 | Translation of the Aw112010 non-canonical ORF encoded protein is required for IL-12 production. **a**, BMDMs were pretreated with cytochalasin D (CD) ($10 \mu\text{M}$) for 1 h, LPS (10 ng ml^{-1}) for 6 h, or non-treated. pHrodo BioParticles were administered for 1 h and cells were assessed for CD11b and pHrodo expression by flow cytometry. Plots are representative of three independent experiments. **b**, BMDMs were infected with *S. Typhimurium* for 6 h. Cells were lysed and CFUs were determined. **c**, BMDMs were pretreated with LPS (100 ng ml^{-1}) for 5 h and infected with *S. Typhimurium* for 1 h, and the release of lactate dehydrogenase (LDH) was measured. RU, relative units. **d**, BMDMs were stimulated with LPS (10 ng ml^{-1}), and *I/12b* expression was determined by qPCR. **e**, Mice were administered PBS ($n = 5$) or LPS ($n = 6$, WT and Aw112010^{Stop}) (10 mg kg^{-1}) for 6 h via intraperitoneal injection. Serum levels of IL-12p40 were determined by ELISA. **f**, BMDMs were stimulated with LPS (10 ng ml^{-1}), and Aw112010 expression was determined by qPCR. **g**, BMDMs were treated with cycloheximide (CHX;

$50 \mu\text{g ml}^{-1}$) or nonsense-mediated decay inhibitor (iNMD; $50 \mu\text{M}$) for 6 h, and Aw112010 expression was determined by qPCR. Fold change was determined relative to vehicle samples. **h**, Predicted minimal free energy of RNA folding of wild-type and mutant (Mut) Aw112010 mRNA. **i**, **j**, BMDMs were subjected to electroporation with indicated plasmids. BMDMs were stimulated with LPS (10 ng ml^{-1}) for 6 h. **i**, Western blot conducted for Aw112010-Flag and β -tubulin. **j**, *I/12b* mRNA was determined by qPCR. Error bars denote s.e.m. Data in **b** and **c** are from three independent experiments conducted with three biological and three technical replicates. Data in **d** are from three biological replicates, and fold change in expression was calculated relative to the non-treated wild-type sample. Data in **f** and **g** are from four independent experiments. Data in **i** and **j** are from three biological replicates. Fold expression in **j** is calculated from a single wild-type EV NT replicate for wild-type cells, and a single Aw112010^{Stop} EV NT replicate for Aw112010^{Stop} cells. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$, unpaired two-tailed *t*-test.

littermate mice for 5 days. Aw112010^{Stop} mice were significantly protected from colitis as measured by weight loss and colonic shortening compared to wild-type counterparts (Fig. 3h, i). Together, these data demonstrate that the translation of the non-canonical ORF in Aw112010 is required for the mucosal inflammatory response.

We next sought to determine the mechanism by which Aw112010 mediates its anti-bacterial and pro-inflammatory function in BMDMs. Wild-type and Aw112010^{Stop} BMDMs showed no difference in their ability to phagocytose or initiate phagosome acidification of pH-sensitive bacterial BioParticles (Fig. 4a). Similarly, intracellular killing or survival of *S. Typhimurium* was also comparable between wild-type and Aw112010^{Stop} BMDMs (Fig. 4b). Furthermore, the ability of wild-type and Aw112010^{Stop} cells to undergo the inflammatory cell death pathway, pyroptosis, was unaltered in Aw112010^{Stop} macrophages compared to wild-type counterparts (Fig. 4c). We next investigated whether Aw112010 was essential for the production of known cytokines responsible for anti-*Salmonella* defence and that contributed physiologically to the intestinal inflammation and inflammatory bowel disease. Notably, although wild-type BMDMs were able to generate a robust LPS induction of IL-12p40 and IL-6, Aw112010^{Stop} macrophages showed a significant deficiency in their production, whereas the release of anti-inflammatory IL-10 was unaltered in Aw112010^{Stop} macrophages (Fig. 4d and Extended Data Fig. 8a–c). To confirm this in vivo, wild-type and Aw112010^{Stop} mice were administered LPS by intraperitoneal injection and serum was collected after 6 h. Again, wild-type animals were able to induce a robust IL-12p40 and IL-6 response that was significantly curtailed in Aw112010^{Stop} mice (Fig. 4e and Extended Data Fig. 8d). Notably, patients with deficiencies in IL-12R are characterized by severe and recurrent *Salmonella* infections, and mice deficient in the IL-12p40 cytokine subunit are also susceptible to *S. Typhimurium* challenge^{24,25}. Furthermore, IL-12p40 has a crucial role in inflammatory

bowel disease, and a neutralizing monoclonal anti-IL12p40 antibody has been shown to be an efficacious treatment in patients with Crohn's disease and in experimental models of colitis^{26,27}. The introduction of a stop codon into Aw112010 causes a major defect in IL-12p40 production and in the ability of mice to combat *Salmonella* infection and undergo mucosal inflammation. However, as premature stop codon introduction into a protein-coding gene can lead to nonsense-mediated decay²⁸, we investigated this phenomenon in Aw112010^{Stop} macrophages. Notably, premature stop codon insertion into Aw112010 does trigger nonsense-mediated decay that can be rescued with the administration of cycloheximide or a specific nonsense-mediated decay inhibitor²⁹ (Fig. 4f, g). Although this demonstrates the vital importance of translation for Aw112010 gene expression, it presents a problem in that we cannot distinguish a potential lncRNA function from a potential protein function. To this end, we generated an expression plasmid that contained the wild-type Aw112010 ORF transcript and an extensively mutated transcript with synonymous nucleotide substitutions in all codons except the CTG start codon. As expected, these RNAs display a very different folding behaviour, but they generate the same protein product (Fig. 4h and Extended Data Fig. 9a–c). We could successfully reintroduce Aw112010 protein expression in Aw112010^{Stop} BMDMs with nuclear electroporation (Fig. 4i). The loss of IL-12p40 expression and release of this cytokine were almost completely restored with both the wild-type and mutated Aw112010 expressing rescue plasmids (Fig. 4j and Extended Data Fig. 9d). Together, our data provide clear evidence that the ability of Aw112010 to drive IL-12p40 production is dependent on the non-canonical ORF-encoded protein and does not act as its lncRNA annotation dictates.

We demonstrate that the translation of functional non-canonical ORFs is a crucial event during the innate immune response to infection and inflammation. A considerable fraction of non-coding RNA genes

not only associate with the ribosome, but also undergo active protein translation. Although we have only shown this in mouse macrophages, it is highly likely that the same will hold true for most cells and tissues in other eukaryotes. Notably, we have demonstrated the translation of one of these ORFs, and the protein it encodes is functional and has a crucial role in host defence and inflammatory disease. Further work is ongoing to uncover the roles of the other identified non-canonical ORFs. We propose that a re-evaluation of the human protein-coding genome is required to identify cryptic non-canonical ORF protein products that may have major implications for human health and disease.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0794-7>

Received: 2 March 2018; Accepted: 17 October 2018;

Published online 12 December 2018.

- Couso, J. P. & Patraquim, P. Classification and function of small open reading frames. *Nat. Rev. Mol. Cell Biol.* **18**, 575–589 (2017).
- Kozak, M. Regulation of translation in eukaryotic systems. *Annu. Rev. Cell Biol.* **8**, 197–225 (1992).
- Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154**, 240–251 (2013).
- Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789–802 (2011).
- Hoagland, M. B., Stephenson, M. L., Scott, J. F., Hecht, L. I. & Zamecnik, P. C. A soluble ribonucleic acid intermediate in protein synthesis. *J. Biol. Chem.* **231**, 241–257 (1958).
- Sanz, E. et al. Cell-type-specific isolation of ribosome-associated mRNA from complex tissues. *Proc. Natl Acad. Sci. USA* **106**, 13939–13944 (2009).
- Clausen, B. E., Burkhardt, C., Reith, W., Renkawitz, R. & Förster, I. Conditional gene targeting in macrophages and granulocytes using LysMcre mice. *Transgenic Res.* **8**, 265–277 (1999).
- Mudge, J. M. & Harrow, J. Creating reference gene annotation for the mouse C57BL6/J genome assembly. *Mamm. Genome* **26**, 366–378 (2015).
- Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130–D135 (2012).
- Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–D73 (2014).
- Carpenter, S. et al. A long noncoding RNA mediates both activation and repression of immune response genes. *Science* **341**, 789–792 (2013).
- Kotzin, J. J. et al. The long non-coding RNA Morrbid regulates Bim and short-lived myeloid cell lifespan. *Nature* **537**, 239–243 (2016).
- Osuna, B. A., Howard, C. J., Kc, S., Frost, A. & Weinberg, D. E. *In vitro* analysis of RQC activities provides insights into the mechanism and function of CAT tailing. *eLife* **6**, e27949 (2017).
- Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
- Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* **4**, e08890 (2015).
- Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
- Kondo, T. et al. Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila* embryogenesis. *Science* **329**, 336–339 (2010).
- Kearse, M. G. & Wilusz, J. E. Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes Dev.* **31**, 1717–1731 (2017).
- Stothard, P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* **28**, 1102, 1104 (2000).
- Wang, H., Wang, Y., Xie, S., Liu, Y. & Xie, Z. Global and cell-type specific properties of lincRNAs with ribosome occupancy. *Nucleic Acids Res.* **45**, 2786–2796 (2017).
- Xiao, Z. et al. De novo annotation and characterization of the translome with ribosome profiling data. *Nucleic Acids Res.* **46**, e61–e61 (2018).
- D'Lima, N. G. et al. A human microprotein that interacts with the mRNA decapping complex. *Nat. Chem. Biol.* **13**, 174–180 (2017).
- de Jong, R. et al. Severe mycobacterial and *Salmonella* infections in interleukin-12 receptor-deficient patients. *Science* **280**, 1435–1438 (1998).
- Lehmann, J. et al. IL-12p40-dependent agonistic effects on the development of protective innate and adaptive immunity against *Salmonella enteritidis*. *J. Immunol.* **167**, 5304–5315 (2001).
- Mannon, P. J. et al. Anti-interleukin-12 antibody for active Crohn's disease. *N. Engl. J. Med.* **351**, 2069–2079 (2004).
- Neurath, M. F., Fuss, I., Kelsall, B. L., Stüber, E. & Strober, W. Antibodies to interleukin 12 abrogate established experimental colitis in mice. *J. Exp. Med.* **182**, 1281–1290 (1995).
- Pulak, R. & Anderson, P. mRNA surveillance by the *Caenorhabditis elegans* smg genes. *Genes Dev.* **7**, 1885–1897 (1993).
- Martin, L. et al. Identification and characterization of small molecules that inhibit nonsense-mediated RNA decay and suppress nonsense p53 mutations. *Cancer Res.* **74**, 3104–3113 (2014).

Acknowledgements We thank J. Alderman, C. Lieber, C. Hughes, L. Evangelisti, E. Hughes-Picard, E. Ryke, L. Machado and C. Castaldi for help in facilitating this work. We thank J. Galan and H. Sun for providing the S. Typhimurium and for discussion on the infection model. We would thank R. Nowarski, M. Healy, K. Baker and N. Palm for comments and discussion on the manuscript. This work was supported by the Howard Hughes Medical Institute and the Blavatnik Family Foundation (R.A.F.). This work was supported in part by the Searle Scholars Program, the Leukemia Research Foundation, an American Cancer Society Institutional Research Grant Individual Award for New Investigators (IRG-58-012-57), the NIH (R01GM122984), and Yale University West Campus start-up funds (to S.S.). A.K. was in part supported by an NIH Predoctoral Training Grant (5T32GM06754 3-12) (S.S.).

Reviewer information Nature thanks M. Raffatellu and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions R.J. conceived the project, performed experiments, analysed the data and wrote the manuscript. L.K. performed all bioinformatics analysis and aided in writing of the manuscript. A.K. performed all the mass spectrometry experiments, analysed data and aided in writing of the manuscript. W.B. performed experiments and contributed major conceptual insight into the work. A.J. and A.G.Y. participated in experimental design, conducted experiments, analysed data and offered vital conceptual insight. O.M.K., J.R.B., M.H.S. and C.D. performed experiments and analysed data. C.C.D.H. helped with bioinformatics analysis and provided conceptual discussion. L.C., P.B., A.G.S. and H.R.S. helped with experiments. S.S. supervised all mass spectrometry work and contributed to the overall interpretation of this work. R.A.F. supervised the project, helped interpret the work and supervised writing of the manuscript.

Competing interests R.A.F. is a scientific advisor to GlaxoSmithKline. All other authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0794-7>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0794-7>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to R.A.F.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Mice. RiboTag (B6N.129 strain) and LysM-Cre (B6.129P2 strain) mice have previously been described and were obtained from Jackson Laboratories^{6,7}. Crossing these mice facilitates the inclusion of a HA-epitope tag on the ribosomal protein RPL22 in all LysM-expressing cells, such as BMDMs and colonic macrophages. Aw112010^{HA} and Aw112010^{Stop} mice were generated as previously described with CRISPR-Cas9 technology³⁰ into C57/B6N embryos. To generate Aw112010^{HA} mice, a ssDNA donor oligonucleotide containing a flexible linker sequence followed by an HA epitope tag sequence flanked with homology arms was provided to facilitate homology-directed repair insertion of the sequence into the C terminus of the Aw112010 ORF. In brief, a guide RNA (**AGAAGGAAGAG GACTTATTGTTT** TAGAGCTAGAAATAGCAAGTTAAATAAGGCTAGT CCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTT) (bold sequence represents guide RNA sequence targeting gene) and ssDNA Ultramer (IDT) donor template (AACCTCAAGTGGAAAAAGCCACCCACTGGGTCGT TCAGGAGAGATCCAGTCTTTAAAGAAGCAAAACggtggttctggtggttctggtggtggttaccatagatttcagattacgctTAGAGAGCAAAATAAGTCCTCTTCTCTCTAGATGTGCATCATCTGCTTCTTCTCTCCCTAGAAGACT) (lowercase sequence represents the knock-in cassette; uppercase sequence represents homology arms) were designed targeting the exon 3 stop codon locus of Aw112010 to generate Aw112010^{HA} mice. To generate Aw112010^{Stop} knock-in mice, CRISPR-Cas9 technology was used to introduce a double-stranded DNA in the second exon of the transcript, in area of high ribosomal protection. A ssDNA donor oligonucleotide containing a small 14-nucleotide sequence (TAATTAATTAATTA) sequence flanked with homology arms was provided to facilitate homology-directed repair insertion into the ORF. This sequence contains a stop codon in all three frames and owing to its even number of nucleotides, it will also frameshift any protein-coding sequence upon insertion. A guide RNA (**CTGCCTGATGCAACAATACCGT TTTAGAGCTAGAAATAGCAAGTTAAATAAGGCTAGTCCGTTATCAACT TGAAAAAGTGGCACCGAGTCGGTGCTTTTTT**) (bold sequence represents guide RNA sequence targeting gene) targeting exon 2 of Aw112010 was designed and co-injected into fertilized C57/B6N eggs with a ssDNA Ultramer (IDT) donor template (TCCTATTCATCTGATCTGCTTCCAGATCCCTCTGATATTATC TTTGGTGGTGTGCTCATCATCTGCCTGATGCAATAaataaataaCAATACCTG GCGTATAAGTCTCTAAGAACGTCGTTAAAGTCTTCTGCCATCAAGCC AAGTAGTGTAGTGTGGG) (lowercase sequence represents the knock-in cassette; uppercase sequence represents homology arms) containing a frameshifting stop insertion and 2 homology arms to allow for homology-directed-repair-mediated genomic integration to generate A112010^{Stop} codon knock-in mice. Single heterozygous founder mice were generated and backcrossed to C57/B6N mice. Experimental groups of wild-type and Aw112010^{Stop} knock-in mice were generated by heterozygote-by-heterozygote breeding. All experiments were performed using littermate control, cohoused mice. All mice experimentation was performed in compliance with Yale Institutional Animal Care and Use Committee protocols. No formal blinding or randomization was conducted; however, control and treated groups were chosen arbitrarily for each experiment. Samples sizes were chosen in line with previous experimental experience and consistent with the broader literature. Mice weights and CFUs were measured in a blinded manner.

RiboTag^{LysM} macrophage RNA isolation and processing. BMDMs were generated from progenitor cells isolated from the femurs and tibias of RiboTag^{LysM} mice and maintained in macrophage-colony stimulating factor (50 ng ml⁻¹) for 7 days. Cells were stimulated with LPS (serotype O111:B4) at the indicated concentrations and times. BMDMs were infected in antibiotic free media with 1 MOI of *S. Typhimurium* for 1 h. Cells were then treated with gentamycin (100 µg ml⁻¹) to kill extracellular bacteria and incubated for a further 5 h before collection. Cells were washed in ice-cold PBS twice, and then RiboTag lysis buffer (containing cycloheximide, heparin and the RNase inhibitor SuperscriptIN) was added directly to the cells on ice as previously described^{6,31–33}. Cell lysates were passed through a 26-gauge needle 10 times and incubated for 30 min on ice to ensure complete lysis. For intestinal macrophage ribosome isolation, control and *S. Typhimurium* infected mice were fasted for 4 h and administered streptomycin (20 mg per mouse) by oral gavage. After 20 h, mice were again fasted for 4 h and gavaged with (2×10^8 CFUs of *S. Typhimurium*). After 24 h, mice were euthanized and the colons were removed. After washing and flushing with PBS, the colon was separated into five equal-sized samples and placed into 1 ml of RiboTag lysis buffer each. Tissue was lysed in a three-step manner. First, tissue was mechanically disrupted with an electronic tissue homogenizer for 30 s. Then homogenized tissue was further processed in a Dounce homogenizer with 10 strokes. Finally, colon lysates were passed through a 26-gauge needle 10 times and incubated on ice for 30 min. Ribosome-RNA-containing supernatants were clarified by centrifugation at 12,000g for 10 min at 4°C. Haemagglutinin-conjugated magnetic beads (Pierce) were added to samples and incubated overnight under gentle inversion at 4°C. Beads were washed three times for 10 min with gentle rotation in high-salt buffer containing cycloheximide. RNA was eluted from HA-beads using Qiagen RLT

buffer containing 2-Mercaptoethanol and anti-foaming DX reagent (Qiagen) by 30 s vortex pulsing. RNA was isolated using an RNeasy micro kit. RNA was sent for sequencing or converted to cDNA using Maxima Reverse transcriptase kit (Thermo). qPCR using Sigma KiCqStart predesigned SYBR green primers was conducted. mRNA for RNA-seq analysis was purified using poly(A)⁺ selection and processed by the Yale Centre for Genome analysis using standard methodology. RNA was sequenced on a HiSeq2000 with 75-bp paired-ended reads.

RiboTag RNA-seq analysis. Fastq files from RNA-seq and RiboTag RNA-seq were aligned to the Ensemble GRCm38.p5 genome using Tophat2 version 2.1.1, and using a gene annotation file which combined all RefSeq, UCSC, Ensembl, Gencode and mirBase annotations for full genome coverage. Remaining missing annotations were added manually from MGI. Cufflinks version 2.2.1 was used for differential analysis. A cut-off of FPKM ≥ 0.1 was used in the RNA-seq to define a 'detectable' lncRNA, and a cut-off of FPKM ≥ 1 in at least one of the three riboTag-RNA-seq treatments was defined as an 'expressed' lncRNA. featureCounts of the Subread package was used to determine reads per feature in the genome³⁴, which was then used with DESeq2 and ggplot2 (<http://ggplot2.tidyverse.org/index.html>) to produce volcano plots³⁵. Rcircos was used to make a circos plot showing the RiboTag RNA-seq data³⁶, using a differential cut-off of 0.01. Only reads which aligned to a single genome locus were used.

Macrophage ribosome profiling and analysis. BMDMs were generated and plated at 1×10^7 cells per treatment group. Non-treated and LPS stimulated macrophages were treated with cycloheximide (50 µg ml⁻¹) for 2 min at 37°C. Cells were then washed with ice-cold PBS containing cycloheximide (50 µg ml⁻¹). Ribosome profiling was conducted using the illumina TruSeq Ribo Profile (mammalian) Kit as per manufacturer's instructions. Prepared libraries were sequenced with a HiSeq 2000 with 75-bp single end reads. In tandem, total RNA from paired samples were subjected to Ribosomal RNA removal and convention RNA-seq. Reads from ribosome profiling experiments had their adapters trimmed using the FASTX-Toolkit version 0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Reads were aligned to Ensemble GRCm38.p5 genome using Tophat2. The accepted hits were filtered using Samtools version 1.5 (<http://samtools.sourceforge.net/>) along the criteria of having an alignment score of 0 to -2, and the number of possible alignments in the genome being less than or equal to 2 to reduce false alignments to repeat regions. Cufflinks was then used for differential expression. DESeq2 and ggplot2 were used for making volcano plots. Sushi and ggplot2 were used to build RNA-seq and ribosome profiling coverage plots for certain features³⁷. A pipeline to determine the translation efficiency and RRS included finding the number of reads covering each feature using featureCounts. Protein coding gene annotations and their UTRs were extracted from our larger GTF annotation file. mRNA sequences from lncRNA exons were extracted using Bedtools version 2.26.0³⁸. To identify non-canonical ORFs in lncRNAs, we first catalogued all transcripts that harboured high quality and unique ribosome profiling reads mapping to exons in the annotated lncRNA family. We then conducted a custom ORFfinder search, in which we relaxed ORF identification parameters pertaining to ORF size and start codon utilization. NCBI ORFfinder and the Sequence Manipulation Suite ORF Finder tools were used to find all ORFs more than 30 nucleotides long using start codons ATG, CTG, TTG and GTG²⁰. Found ORFs were searched using BLASTX+ version 2.6.0 against lncRNA sequences that included their introns to find the final position of ORFs within the lncRNAs. UTRs were defined as 50 nucleotide bases directly following the coding sequence. Translational efficiency and RRS were calculated according as previously described³. Translational efficiency was the number of ribosomal reads covering the exons divided by the number of RNA reads in the exon, normalized for length. In brief, RRS interrogates the known phenomenon of ribosome release from a transcript directly after recognition of the ORFs cognate stop codon³⁹ by enumerating the number of ribosome profiling reads before and after the in frame stop codon of a given mRNA. RRS is the result of ribosomal reads covering the exon divided by RNA reads covering the exon, divided by the result of ribosome reads covering the 3' UTR divided by RNA reads covering the 3' UTR. $RRS \geq 7$ and translational efficiency ≥ 0.0001 values are considered that of the annotated protein coding genome. Exons had to have at least 1 ribosomal read for these metrics to be calculated. A FPKM cut-off of ≥ 4 in either treatment was used when identifying top significant differentially regulated features. Significance of $P < 0.05$ was used. RibORF version 0.1 was used to identify transcripts undergoing active translation measured by the correct alignment of the ribosome A-site, 3-nucleotide periodicity of translating ribosomes and coverage uniformity across all codons of protein coding genes. RibORF was then used to assign PME scores to the ORFs found with ORFfinder. A PME above 0.6 was used to predict positive coding potential. RiboCode version 1.2.10 was used to identify ORFs within the ribosome profiling data using start codons ATG, CTG, TTG and GTG. A combined analysis $P < 0.05$ was used to determine protein-coding potential. The nucleotide sequences of all predicted coding ORFs found were searched for homology in the human CHES 2.0 genome (<http://ccb.jhu.edu/chess/>) using BLASTX+ version 2.6.0. PhyloCSF was used on the top hit of ORFs for which homology was found

to further assess coding potential⁴⁰. All lncRNA RNA sequencing and analysis is provided in Supplementary Table 1.

Expression of non-canonical ORFs. Gene blocks corresponding to the open readings identified by ribosomal profiling were synthesized by IDT and inserted into pCMV6-entry tagged cloning vector (PS100001) using MLU I and Sgf I restriction digestion. Plasmids were verified by Sanger sequencing and propagated by transformation in TOP10 competent cells. HEK293 cells were transfected using Transit LT1 liposomal transfection reagent (Mirus) as per manufacturer's instructions. For confocal microscopy studies, HEK293 (obtained directly from ATCC) cells were transfected with 0.5 µg of plasmid DNA and incubated at 37 °C for 24 h. Cells were fixed in methanol. Anti-Flag monoclonal antibody (M2 clone), phalloidin 647 (Santa Cruz) and DAPI (Sigma) were used. Confocal imaging was conducted with a Nikon-Ti microscope combined with UltraVox spinning disk (PerkinElmer) and data was analysed using the Volocity software (PerkinElmer). Cells were not tested for mycoplasma and were not further authenticated.

Aw112010^{HA} immunoblotting and immunoprecipitation. Wild-type and Aw112010^{HA} BMDMs were generated and treated with LPS (10 ng ml⁻¹) for the indicated times. Protein lysates and haemagglutinin immunoprecipitation fractions were generated using the Pierce HA-Tag Magnetic IP/Co-IP Kit as per manufacturer's instructions. Protein lysates and haemagglutinin immunoprecipitation samples were resolved on NuPAGE 4–12% Bis-Tris protein gels using MES running buffer (Invitrogen) and transferred to nitrocellulose membrane. Detection of endogenous Aw112010 was conducted using a monoclonal anti-HA antibody (HA1.1 clone) and β-tubulin (E7 clone) was used as a loading control.

Proteomics methods. BMDMs (1 × 10⁷) stimulated with LPS (10 ng ml⁻¹) for 6 h were lysed and subjected to haemagglutinin immunoprecipitation as described above. An aliquot of haemagglutinin-purified protein lysate was boiled for 15 min in 1% SDS followed by chloroform-methanol precipitation⁴¹. Reduction, alkylation, and Lys-C protease digestion were performed according to standard protocols^{42,43}. All resulting peptides were purified and desalted using a SepPak Classic SPE cartridge (Waters) according to the manufacturer's instructions and dried in a rotary vacuum centrifuge. Samples were resuspended in 0.1% trifluoroacetic acid and diluted to approximately 0.5 µg µl⁻¹. Peptide standards (JPT Peptide Technologies) with ¹³C₆¹⁵N₄ R isotopic labels were added to a 100 ng ml⁻¹ final concentration⁴⁴. A total volume of 5 µl of sample was injected onto an analytical column (75 µm × 50 cm PicoFrit column packed with 1.9 µm ReproSil-Pur 120Å C18-AQ resin) using ACQUITY UPLC M-Class (Waters) and a Q Exactive Plus (Thermo). Separation was performed on a 330 min nonlinear gradient from 1% mobile phase B to 99% mobile phase B (mobile phase A: 1% acetonitrile 0.1% formic acid in water, mobile phase B: 80% acetonitrile 0.1% formic acid in water); MS: 70,000 resolution, 3 × 10⁶ AGC target, 300–1,700 m/z scan range; dd-MS2: top10 method, 17,500 resolution, 1 × 10⁶ AGC target, 10 loop count, 1.6 m/z isolation window, 27 normalized collision energy. In all experiments, a full mass spectrum was followed by ten parallel reaction-monitoring scans at 17,500 resolution 2 × 10⁵ AGC target, 4 m/z isolation window 100 ms maximum injection time) as triggered by an inclusion list. ProteoWizard MS Convert was used for peak picking and files were analysed using Mascot (version 2.5.1). RNA-seq data from LPS and non-treated mouse macrophages were aligned to Ensemble GRCm38.p5 genome and translated in three reading frames using CLC SequenceViewer (Qiagen) and the resulting databases as well as a contaminant database were used for proteomics searches. Carbamidomethyl (C) was set as a fixed modification, and ¹³C₆¹⁵N₄ R, oxidation (M), and acetyl (N-term) as variable modifications. The false discovery rate was set to 1%.

S. Typhimurium infection. Before infection, 8–10-week-old mice were restricted from food and water for 4 h followed by gavage of streptomycin (20 mg). After 20 h, mice were fasted again for 4 h and infected with streptomycin resistant *Salmonella enterica* subsp. *enterica* serovar Typhimurium (SL1344 strain, provided by J. Galan). *S. Typhimurium* was maintained as a glycerol stock at -80 °C. Before infection, bacteria were propagated overnight in LB containing streptomycin (100 µg ml⁻¹). Bacteria was subcultured for 4 h the next day in antibiotic-free LB containing 0.3 M NaCl to return it to log phase growth and increase virulence⁴⁵. Using spectrophotometry, bacterial CFU was calculated with an infection dose ranging from 1 × 10² to 1 × 10³ CFUs per mouse. To calculate faecal CFU, faecal pellets were resuspended in PBS at 50 mg ml⁻¹ and vortexed for 20 min. Bacteria containing supernatants were clarified by centrifugation at 50g for 10 min. Serial dilutions were conducted, and bacteria plated in triplicate on LB streptomycin (100 µg ml⁻¹) plates. For caecal, liver and spleen CFU enumeration, organs were isolated and weighed, and added to 2 ml of PBS. Tissue was dissociated with GentleMacs C Tubes (Miltenyi Biotec) as per manufacturer's instructions. CFU counts were calculated using similar methodology as above. All CFU counts were preformed blinded. Mice that lost 30% body weight, or that were unresponsive, were euthanized.

Confocal microscopy for splenic *Salmonella* dissemination. Wild-type and Aw112010^{Stop} codon mice were infected with 1 × 10² CFUs as described above.

Mice were euthanized 3 days after infection. For in situ immunofluorescence, spleens were dissected, fixed in 4% paraformaldehyde for 1 h at 4 °C, followed by incubation in 10% sucrose in PBS for 1 h, 20% sucrose in PBS for 1 h, and 30% sucrose in PBS overnight at 4 °C, all under gentle agitation. Tissue samples were frozen in OCT on dry ice and kept at -80 °C until sectioning. Sections 8–10-µm thick were prepared with a cryostat (Leica). Spleen sections were permeabilized with Perm/Wash buffer (BD Biosciences) for 10 min and blocked with Protein Block (Dako) for 7 min. Primary antibodies included anti-F4/80 (BM8), anti-B220 (RA3-6B2) and anti-Salmonella (Abcam ab35156). Primary and secondary antibodies were incubated in Perm/Wash buffer for 1 h. After washing with Perm/Wash buffer, sections were mounted with ProLong Gold with DAPI (Invitrogen), covered and sealed with nail polish. Confocal imaging was conducted with a Nikon-Ti microscope combined with UltraVox spinning disk (PerkinElmer) and data was analysed using the Volocity software (PerkinElmer).

DSS colitis induction. DSS colitis was conducted as previously described³⁰. In brief, 10–12-week-old mice were administered 2.5% DSS (MP Bio) for 5 days and returned to regular drinking water and monitored daily. Weight loss was measured every day. On day 12, mice were euthanized and colons were extracted. Colonic shortening was used as a metric of colitis severity.

Phagocytosis assay. Wild-type and Aw112010^{Stop} BMDMs were generated and plated on non-tissue culture-treated non-adherent plates and pretreated with the phagocytosis inhibitor cytochalasin D (10 µM) for 1 h, LPS (10 ng ml⁻¹) for 6 h or left non-treated. Cells were then administered with 1 mg ml⁻¹ pHrodo Red *Escherichia coli* BioParticles (Invitrogen) conjugates for 1 h. Cells were isolated, washed and stained with anti-CD11b (M1/70). CD11b positive cells were assessed for pHrodo positivity which indicates cells which have phagocytosed the pH sensitive bio particles and initiated phagosome acidification using a LSRII flow cytometer.

Intracellular *Salmonella* survival assay. Wild-type and Aw112010^{Stop} BMDM generation and *S. Typhimurium* bacterial culturing was conducted as previously described. On day 7, the BMDM culture media was replaced with antibiotic-free media. BMDMs (1 × 10⁶) were infected in triplicate with 1 × 10⁷ CFUs of *S. Typhimurium* (MOI 10) by centrifugation at 800g at 37 °C for 10 min and returned to the incubator for a further 20 min. Media was then removed and replaced with media containing gentamycin (100 µg ml⁻¹) to kill extracellular bacteria for 1 h. Media was replaced with fresh media including gentamycin (25 µg ml⁻¹) for 4.5 h. Cells were lysed in a 1% Triton and 0.1% SDS buffer for 5 min under gentle agitation. Cell lysates were plated on streptomycin (100 µg ml⁻¹) containing LB plates and CFUs enumerated.

BMDM cell death assay. Wild-type and Aw112010^{Stop} BMDMs were generated and seeded at 50,000 cells per 96 well plate in triplicate. Cells were stimulated with LPS (100 ng ml⁻¹) for 5 h. Media was replaced with fresh antibiotic free media containing LPS (100 ng ml⁻¹) and *S. Typhimurium* (MOI 100). Cells were centrifuged at 800g at 37 °C for 10 min. Cells were returned to the incubator for a further 50 min. Supernatants were collected and clarified. The Pierce LDH Cytotoxicity Assay Kit was used to measure cell death, as per manufacturer's instructions.

In vitro BMDM LPS-induced cytokine measurement. Wild-type and Aw112010^{Stop} BMDMs were generated and stimulated with LPS (10 ng ml⁻¹) for the indicated times. Supernatants were collected and clarified by centrifugation (12,000g, 10 min at 4 °C) and analysed for IL-12p40, IL-6 and IL-10 by ELISA (R&D Duosets). Similarly, BMDMs were generated and stimulated as above and lysed in Trizol and RNA extracted as per manufacturer's instructions. RNA was equalized to 700 ng and converted to cDNA using Maxima Reverse transcriptase kit (Thermo). qPCR was conducted using Sigma KicqStart predesigned SYBR green primers as indicated.

In vivo LPS-induced cytokine measurement. Wild-type and Aw112010^{Stop} 8–10-week-old mice were weighed and administered 100 µl of either PBS or LPS (Serotype O111:B4, Enzo LifeSciences) (10 mg kg⁻¹) via intraperitoneal injection. After 6 h, mice were euthanized and serum collected and analysed for IL-6 and IL-12p40 cytokines by ELISA (R&D Duosets).

Nonsense-mediated decay studies. Wild-type and Aw112010^{Stop} BMDMs were generated and stimulated with cycloheximide (50 µg ml⁻¹) or a nonsense-mediated decay inhibitor (Calbiochem) (50 µM) for 6 h. RNA was extracted and qPCR conducted for Aw112010 mRNA expression.

RNA and protein structural prediction tools. Two Aw112010 rescue expression plasmids were generated. The wild-type Aw112010 mRNA sequence was cloned into the pCMV6 expression vector:

CTGAGCTGCAAGATGTCTCCCATCCCTCTGATATTATCTTTGGTGG
TGTGCTCATCTGCTGCCTGATGCAACAATACCTGGCGTATAAGTCTTCTA
AGAACGTCGTTAAAGTCTTCTGCCATCAAGCCAATGATGTACATATATA
CCAGACCCAGGTGCTCATGACAAACACACTGAAACCTCAAGTGGAAA
AAGCCACCCACTGGGTGCTTTCAGGAGAGATCCAGTCTTTAAAGAAGC
AAAAC.

The mutant *Aw112010* mRNA sequence contains extensive synonymous mutation: CTGTCGTGTAAATGTCACCTATCCCCTAATCTTCATTTTCGGCGGGGTCCTTATTATTGTTTAAATGCAGCAGTATCTAGCATACAAATCATCGAAAAATGTAGTAAAGGTATTTGTCCAGGCGAACGACGTTTCACATTTATCAAACCTCAAGTGGTAATGACGAATCTCTCGAGACAAGCAGCGGGAAGTCACATCCGTTAGGCCGGTCCGGCGAAATACAATCGTTGAAAAAACAGAAT.

The Vienna RNA package was used to calculate minimum free energy structures of each RNA⁴⁶. Both mRNA sequences encode the same protein product: LSCKMSPILIFGGVLIICLMQQLAYKSSKNVVKVFCHQANDVHIYQTQVVM TNTLETSSGKSHPLGRSGEIQSLKKQN. To predict the structure of this protein we used the Quark software package⁴⁷.

Aw112010 rescue experiments. Wild-type and *Aw112010*^{Stop} BMDMs were generated and cultured on non-tissue cultured treated non-adherent plates for 7 days. BMDMs (1×10^6) were electroporated with 1 μ g of rescue expression vector as indicated using the Amaxa Mouse Macrophage Nucleofector Kit using a Nucleofector 2b Device as per manufacturer's instructions (Lonza). Cells were rested for 6 h and stimulated as indicated. Expression of *Aw112010* was confirmed by anti-Flag tag immunoblotting. IL-12p40 protein and *Il12b* mRNA gene expression were measured as previously described.

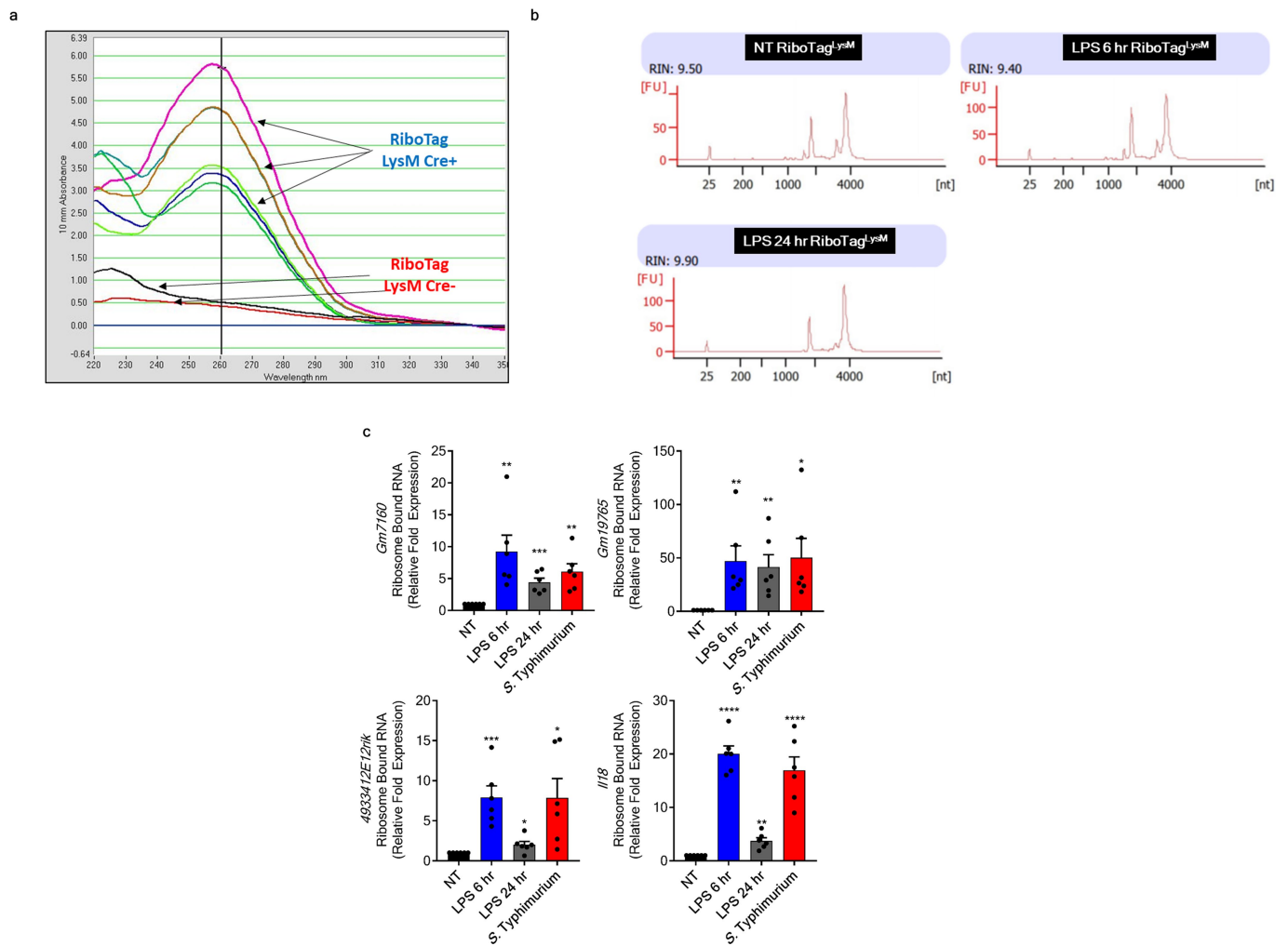
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

RNA-seq, RiboTag RNA-seq, and ribosome profiling data that support the findings of this study have been deposited in the Gene Expression Omnibus (GEO) repository with the accession code GSE120762. All lncRNA RNA-seq, RiboTagSeq, ribosome profiling sequencing and analysis can also be found in Supplementary Table 1.

30. Nowarski, R. et al. Epithelial IL-18 equilibrium controls barrier function in colitis. *Cell* **163**, 1444–1456 (2015).
31. Gabanyi, I. et al. Neuro-immune interactions drive tissue programming in intestinal macrophages. *Cell* **164**, 378–391 (2016).

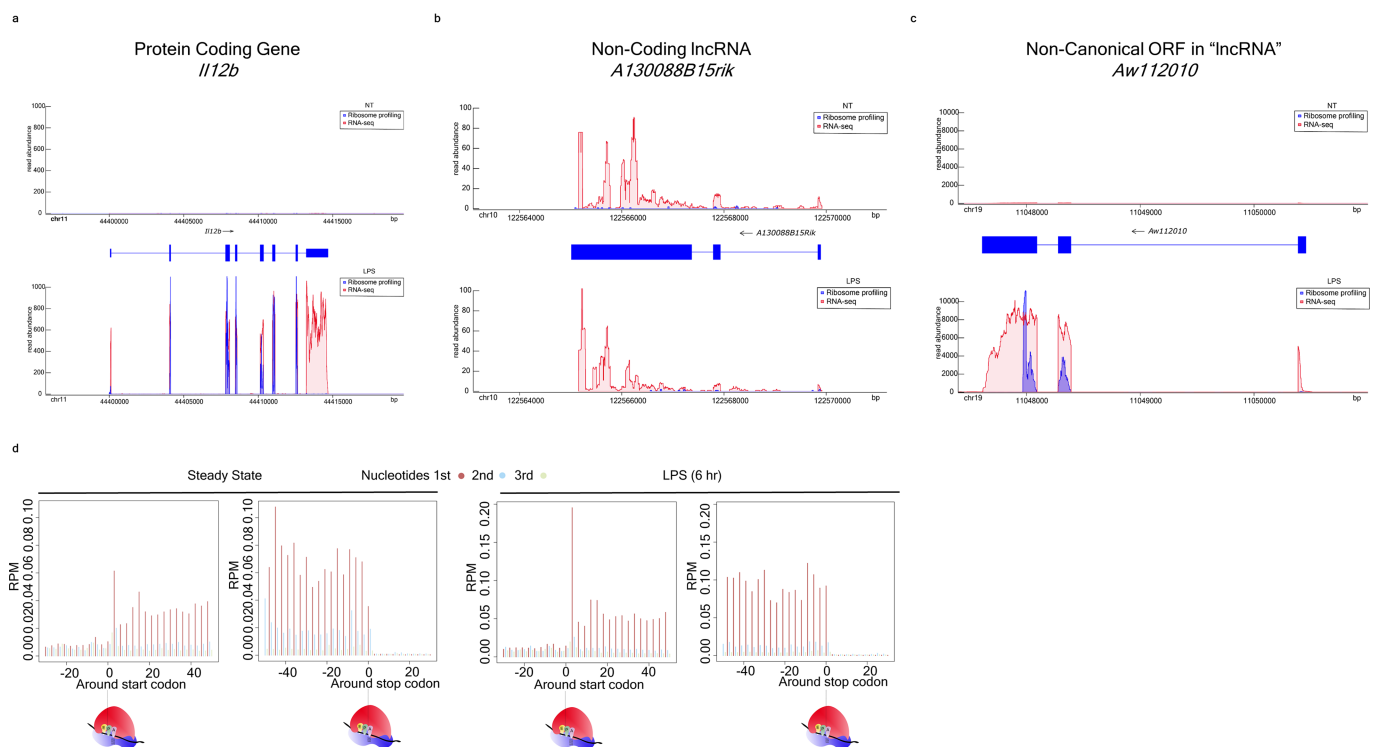
32. Obrig, T. G., Culp, W. J., McKeen, W. L. & Hardesty, B. The mechanism by which cycloheximide and related glutarimide antibiotics inhibit peptide synthesis on reticulocyte ribosomes. *J. Biol. Chem.* **246**, 174–181 (1971).
33. Schneider-Poetsch, T. et al. Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. *Nat. Chem. Biol.* **6**, 209–217 (2010).
34. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
35. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
36. Zhang, H., Meltzer, P. & Davis, S. RCircos: an R package for Circos 2D track plots. *BMC Bioinformatics* **14**, 244 (2013).
37. Phanstiel, D. H., Boyle, A. P., Araya, C. L. & Snyder, M. P. Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics* **30**, 2808–2810 (2014).
38. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
39. Frolova, L. et al. A highly conserved eukaryotic protein family possessing properties of polypeptide chain release factor. *Nature* **372**, 701–703 (1994).
40. Perte, M. et al. Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. Preprint at <https://www.biorxiv.org/content/early/2018/05/28/332825> (2018).
41. Wessel, D. & Flügge, U. I. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* **138**, 141–143 (1984).
42. Slavoff, S. A. et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **9**, 59–64 (2013).
43. Gundry, R. L. et al. Preparation of proteins and peptides for mass spectrometry analysis in a bottom-up proteomics workflow. *Curr. Protoc. Mol. Biol.* **Chapter 10**, Unit10.25 (2009).
44. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W. & Gygi, S. P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl Acad. Sci. USA* **100**, 6940–6945 (2003).
45. Chen, L. M., Kaniga, K. & Galán, J. E. Salmonella spp. are cytotoxic for cultured macrophages. *Mol. Microbiol.* **21**, 1101–1115 (1996).
46. Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R. & Hofacker, I. L. The Vienna RNA websuite. *Nucleic Acids Res.* **36**, W70–W74 (2008).
47. Xu, D. & Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80**, 1715–1735 (2012).



Extended Data Fig. 1 | RiboTag RNA isolation and mRNA expression.

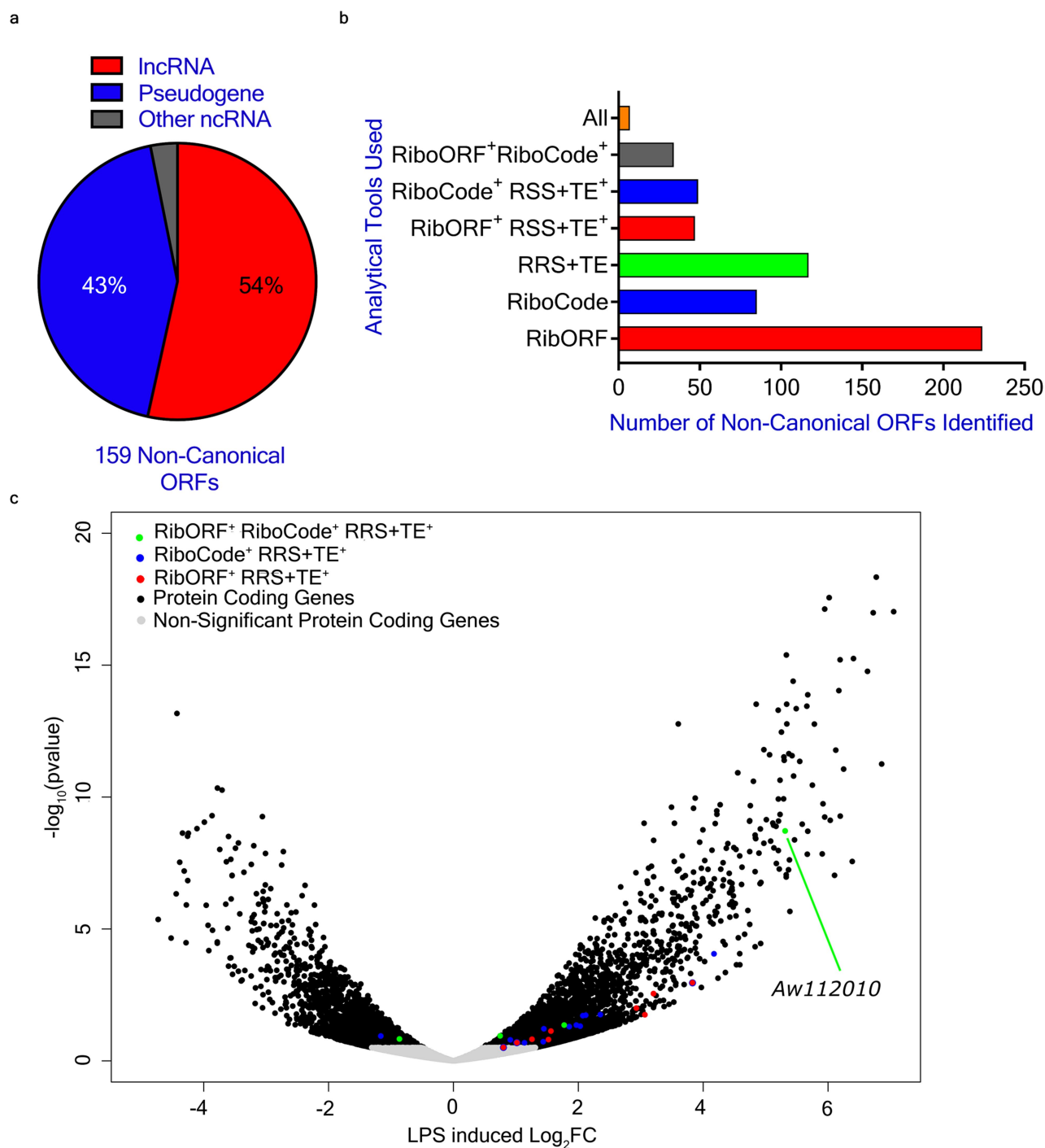
a, Nanodrop analysis of ribosome-associated RNA isolated from RiboTag (Cre⁻) and RiboTag^{LysM} (Cre⁺) mice showing no detected RNA isolated from Cre⁻ BMDMs. **b**, Bioanalyzer (Agilent) traces and RNA integrity number (RIN) of ribosome-associated RNA isolated from BMDMs from

RiboTag^{LysM} mice non-treated or stimulated with LPS for 6 or 24 h. **c**, qPCR analysis of ribosome-associated transcripts from non-treated BMDMs, or BMDMs stimulated with LPS (10 ng ml⁻¹) or infected with *S. Typhimurium* at an MOI of 1 for 6 h. Data are mean ± s.e.m. from six biological replicates.



Extended Data Fig. 2 | Ribosome profiling, RNA-seq read tracing and RibORF analysis. **a–d**, Wild-type BMDMs were non-treated or stimulated with LPS (10 ng ml^{-1}) for 6 h and ribosome profiling was conducted. Data are representative of two biological replicates. **a**, Pattern of RNA-seq transcriptional reads (red) and ribosome profiling translational reads (blue) for *Il12b* from non-treated (top) and LPS-stimulated (bottom) BMDMs. The gene structure of *Il12b* is located in the centre, with a thin blue line representing the introns and wide blue rectangles indicating exonic structure. Thinner exonic structures represent annotated 5' and 3' UTRs. **b**, Pattern of RNA-seq transcriptional reads (red) and ribosome profiling translational reads (blue) for a non-RiboTag-identified lncRNA, *A130088B15rik*, from non-treated (top) and LPS-stimulated (bottom)

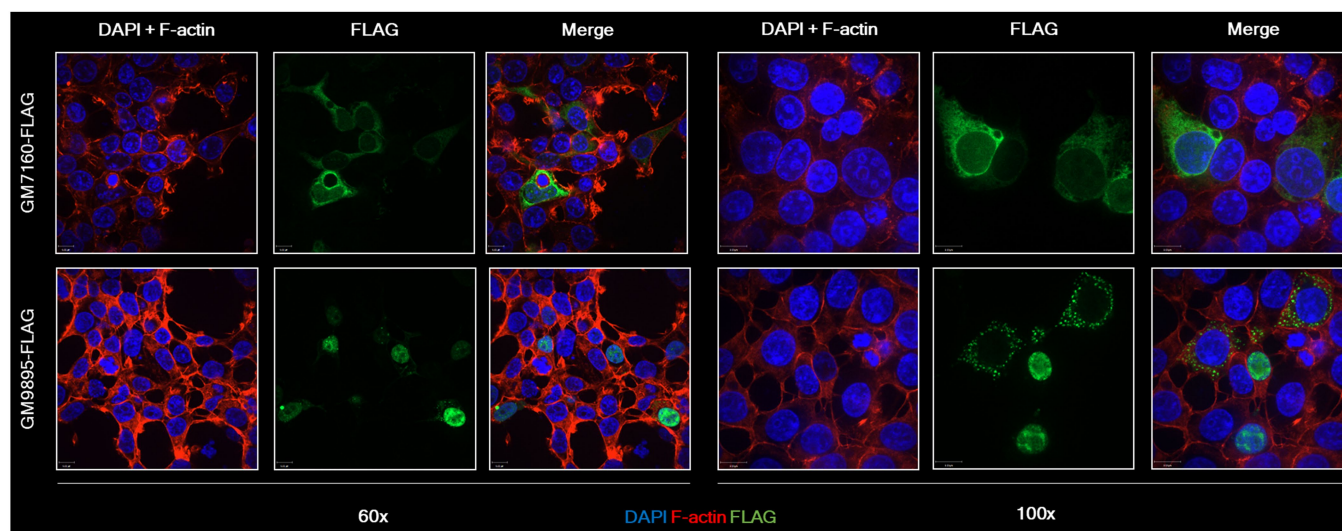
BMDMs. The gene structure of *A130088B15rik* is located in the centre, with a thin blue line representing the introns and wide blue rectangles indicating exonic structure. **c**, Pattern of RNA-seq transcriptional reads (red) and ribosome profiling translational reads (blue) for a RiboTag-identified lncRNA, *Aw112010*, from non-treated (top) and LPS-stimulated (bottom) BMDMs. The gene structure of *Aw112010* is located in the centre, with a thin blue line representing the introns and wide blue rectangles indicating exonic structure. **d**, RibORF analysis of read distribution (reads per million mappable reads; RPM) around start and stop codons of known, annotated protein-coding genes in steady state and LPS-stimulated samples.



Extended Data Fig. 3 | Breakdown of different analytical approaches to predict protein coding lncRNAs. **a**, RiboCode analysis of ribosome-profiling data identifies 85 ORFs within lncRNAs with protein-coding potential. **b**, Comparison of non-canonical ORFs identified by RibORF, RRS and translation efficiency, and RiboCode analytical strategies from BMDMs expressing lncRNA using ribosome profiling. **c**, Wild-type

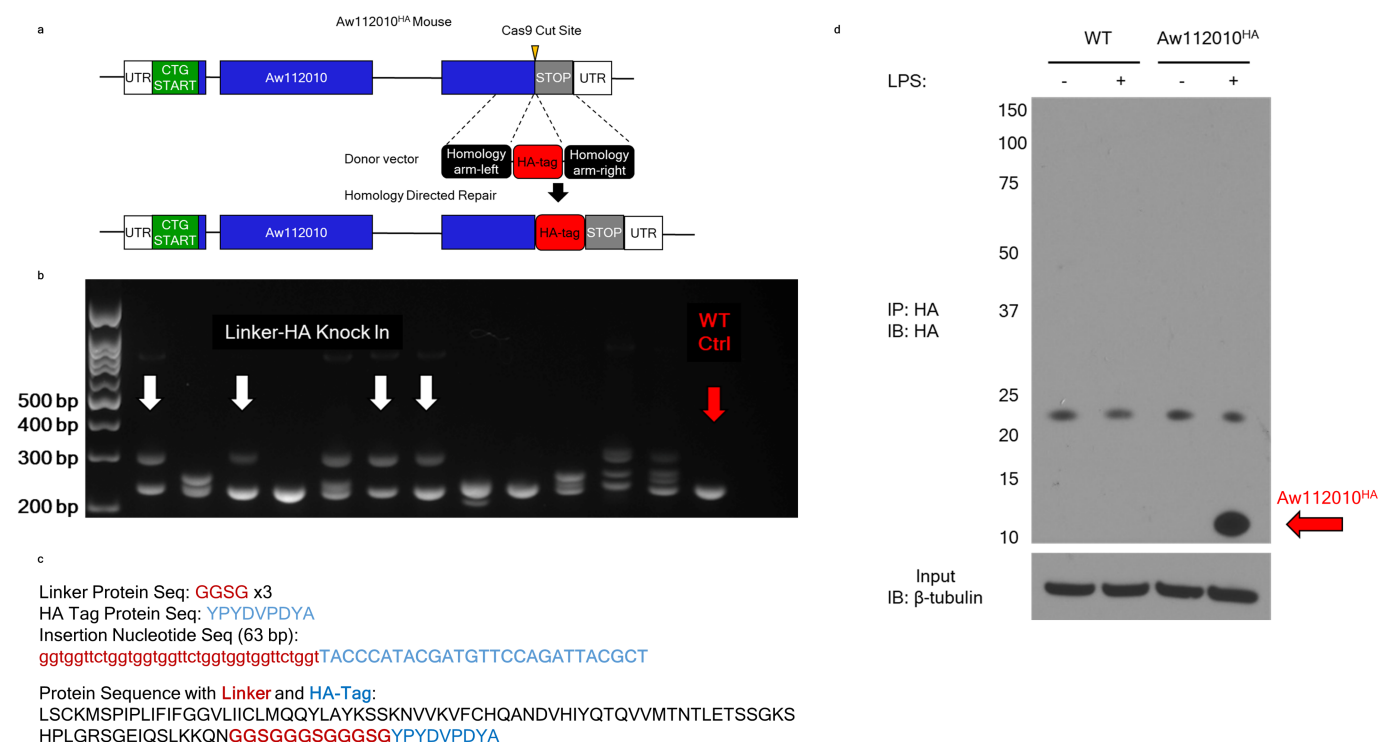
BMDMs were non-treated or stimulated with LPS (10 ng ml⁻¹) for 6 h and ribosome profiling was conducted. Data are representative of two biological replicates. Volcano plot of LPS-induced differentially regulated genes identified by RibORF, RiboCode, RRS and translation efficiency analysis.

a



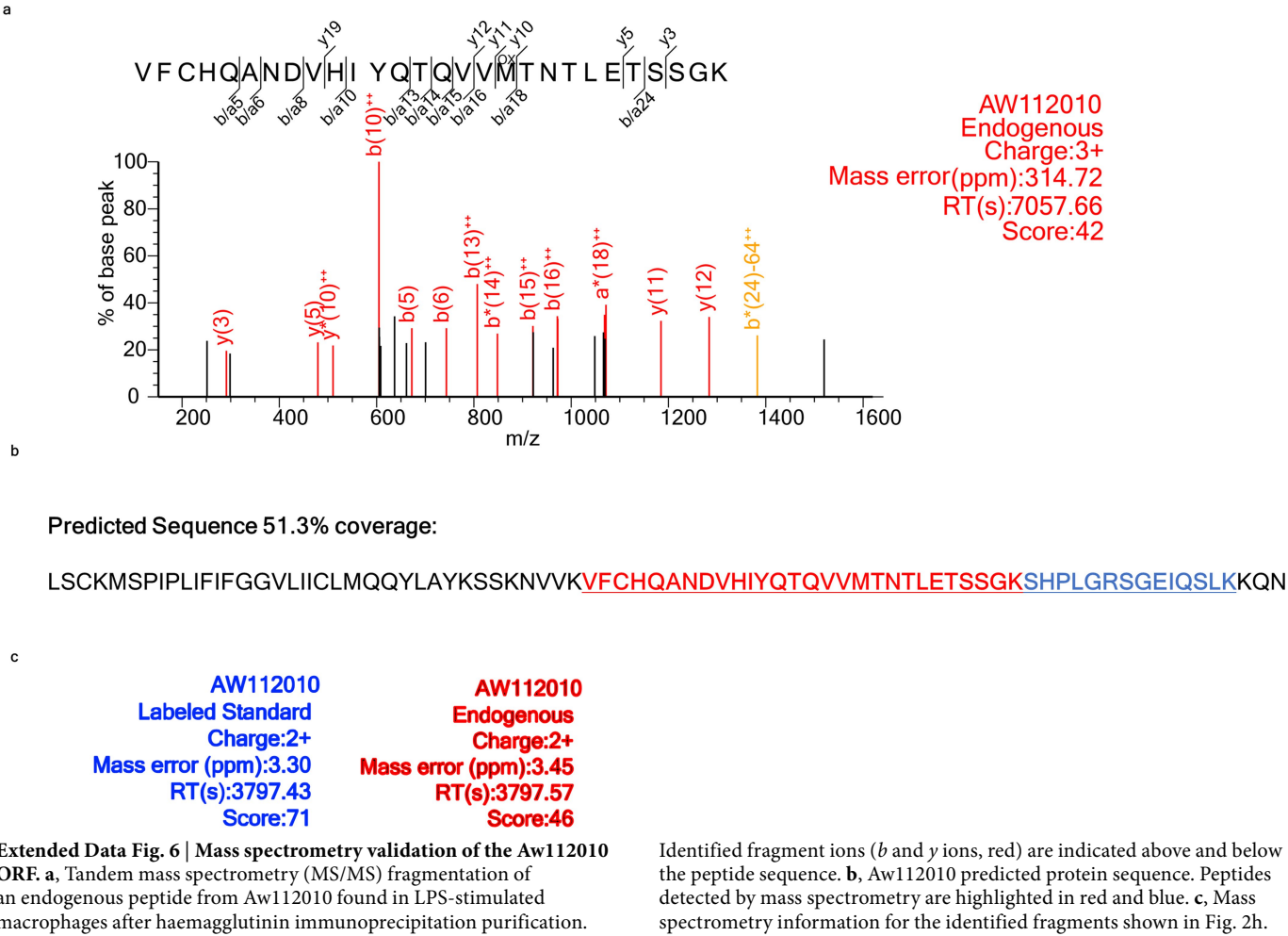
Extended Data Fig. 4 | Overexpression of non-canonical ORFs reveals distinct subcellular localization. **a**, HEK293 cells were transfected with 500 ng of Flag-tagged plasmids encoding the non-canonical ORFs GM7160 and GM9895. Cells were fixed and stained with DAPI (blue,

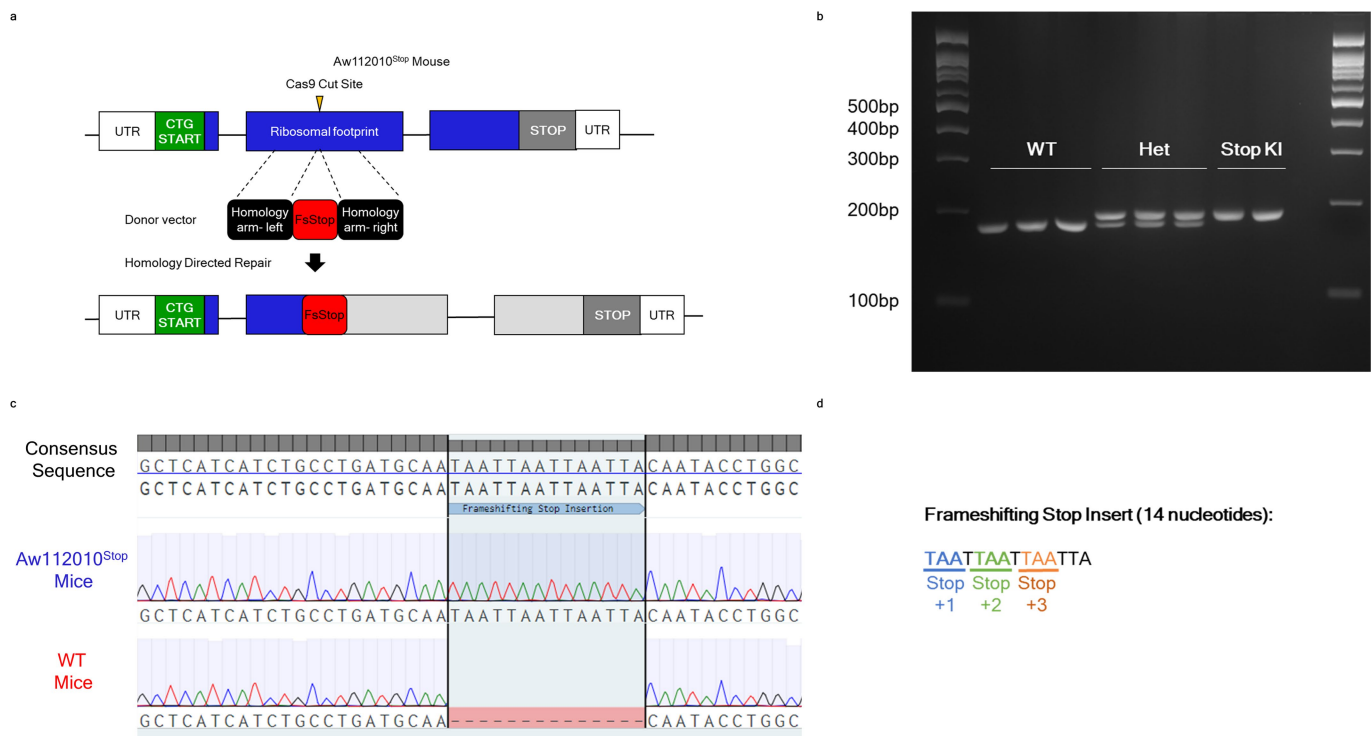
nucleus), phalloidin (red, cytoskeletal F-actin) and anti-Flag (green, ORF of interest). Original magnifications, $\times 60$ and $\times 100$. Data are representative of three or more independent experiments.



Extended Data Fig. 5 | Aw112010^{HA} mouse characterization. **a**, Schematic representation of Aw112010^{HA} knock-in mice. **b**, Genotyping for Aw112010^{HA} mice from CRISPR–Cas9 injections. **c**, Sequence information for GGSG(\times 3)–HA-tag insertion used to generate Aw112010^{HA} mice. **d**, Wild-type and AW112010^{HA} BMDMs were left untreated or stimulated

with LPS (10 ng ml⁻¹) for 6 h. Protein lysates were generated and incubated overnight with anti-haemagglutinin magnetic beads. Purified lysates were probed for haemagglutinin by western blot. Whole-cell lysates were used as a loading control and probed for β -tubulin. Data are representative of three independent experiments.





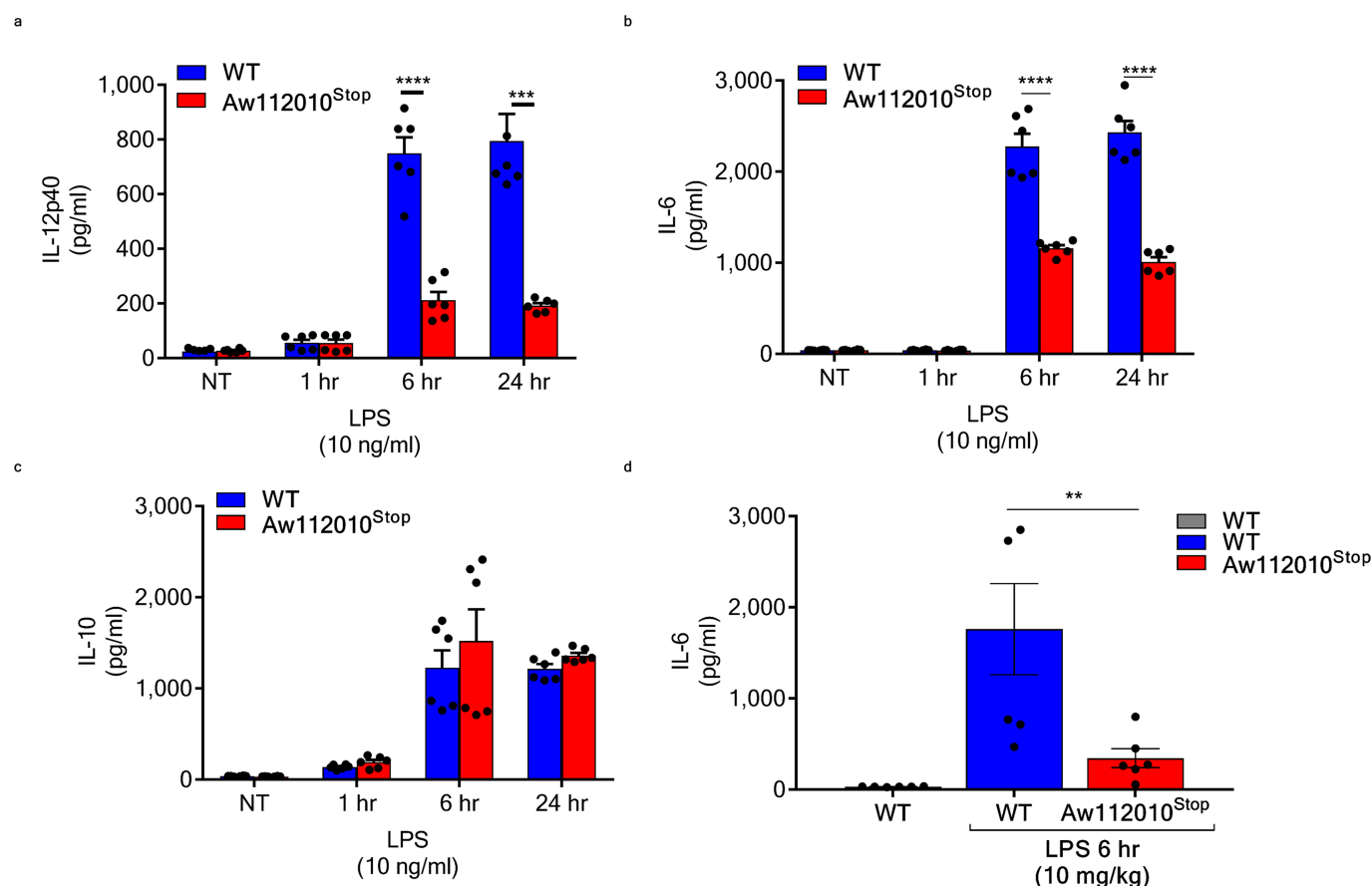
Extended Data Fig. 7 | Characterization of Aw112010^{Stop} mice.

a, Schematic representation of Aw112010^{HA} knock-in (KI) mice.

b, Genotyping for Aw112010^{Stop} mice generated using CRISPR–Cas9.

Het, heterozygous. **c**, Sanger sequencing of the frameshifting stop codon

insertion in Aw112010^{Stop} mice and wild-type controls. **d**, Sequence of frameshifting stop insertion. Stop codons and frame positions are indicated below the sequence.



Extended Data Fig. 8 | Cytokine production in wild-type and Aw112010^{Stop} macrophages and mice. a–c, Wild-type and Aw112010^{Stop} BMDMs were stimulated with LPS for indicated times and supernatants were analysed for IL-12p40, IL-6 and IL-10 by ELISA. Data are from six biological replicates conducted over two independent experiments.

d, Mice were administered PBS ($n = 5$) or LPS ($n = 6$, WT and Aw112010^{Stop}) (10 mg kg^{-1}) for 6 h via intraperitoneal injection. Serum was analysed for IL-6 by ELISA. Error bars denote s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$, unpaired two-tailed t -test.

Design of amidobenzimidazole STING receptor agonists with systemic activity

Joshi M. Ramanjulu^{1,5*}, G. Scott Pesiridis^{1,5}, Jingsong Yang^{2,5}, Nestor Concha⁴, Robert Singhaus¹, Shu-Yun Zhang², Jean-Luc Tran¹, Patrick Moore¹, Stephanie Lehmann³, H. Christian Eberl³, Marcel Muelbauer³, Jessica L. Schneck⁴, Jim Clemens⁴, Michael Adam², John Mehlmann¹, Joseph Romano¹, Angel Morales¹, James Kang¹, Lara Leister¹, Todd L. Graybill¹, Adam K. Charnley¹, Guosen Ye⁴, Neysa Nevins⁴, Kamelia Behnia¹, Amaya I. Wolf¹, Viera Kasparcova¹, Kelvin Nurse⁴, Liping Wang⁴, Yue Li⁴, Michael Klein⁴, Christopher B. Hopson², Jeffrey Guss⁴, Marcus Bantscheff³, Giovanna Bergamini³, Michael A. Reilly¹, Yiqian Lian², Kevin J. Duffy², Jerry Adams², Kevin P. Foley¹, Peter J. Gough¹, Robert W. Marquis¹, James Smothers^{2,6}, Axel Hoos^{2,6} & John Bertin^{1,6}

Stimulator of interferon genes (STING) is a receptor in the endoplasmic reticulum that propagates innate immune sensing of cytosolic pathogen-derived and self DNA¹. The development of compounds that modulate STING has recently been the focus of intense research for the treatment of cancer and infectious diseases and as vaccine adjuvants². To our knowledge, current efforts are focused on the development of modified cyclic dinucleotides that mimic the endogenous STING ligand cGAMP; these have progressed into clinical trials in patients with solid accessible tumours amenable to intratumoral delivery³. Here we report the discovery of a small molecule STING agonist that is not a cyclic dinucleotide and is systemically efficacious for treating tumours in mice. We developed a linking strategy to synergize the effect of two symmetry-related amidobenzimidazole (ABZI)-based compounds to create linked ABZIs (diABZIs) with enhanced binding to STING and cellular function. Intravenous administration of a diABZI STING agonist to immunocompetent mice with established syngeneic colon tumours elicited strong anti-tumour activity, with complete and lasting regression of tumours. Our findings represent a milestone in the rapidly growing field of immune-modifying cancer therapies.

Cancer immunotherapy has transformed the treatment of cancer, with immune checkpoint inhibitors directed against the programmed cell death 1 ligand (PDL-1) demonstrating clinical efficacy for multiple tumour types. The search for additional immune modulators beyond those that directly target the adaptive immune response has been extended to innate immune activation, which is expected to enhance tumour immunogenicity. The cyclic GMP-AMP synthase (cGAS)–STING pathway has emerged as an important intrinsic tumour-sensing mechanism that sets this pathway apart from other innate immune-sensing pathways that have been proposed to enhance tumour immunogenicity, such as the Toll-like receptor (TLR) pathway^{4,5}. Tumour-derived DNA activates cGAS to produce cGAMP, the endogenous ligand of STING, resulting in downstream signalling cascade via recruitment of serine/threonine-protein kinase (TBK1), phosphorylation of the interferon regulatory transcription factor IRF3, and production of type I interferon (IFN), among other proinflammatory cytokines⁴. Type I interferons selectively stimulate cross-presentation of tumour antigens and mobilization of tumour-specific CD8 T cells, which prime the adaptive immune response against tumours^{6,7}. Pharmacological activation of STING via intratumoral delivery with modified cyclic nucleotides leads to potent and durable regression of tumours in syngeneic tumour models⁸. However, synthetic small molecule STING agonists that are human active and suitable for systemic administration have not been reported.

To identify ligands that modulate STING function, we used high-throughput screening of small molecules that compete with the binding of radio-labelled cGAMP to the C-terminal domain (CTD; amino acids (aa) 149–379) of human STING (Extended Data Table 1). This approach identified a series of small-molecule ABZIs that showed modest yet reproducible inhibition of ³H-cGAMP binding to STING (for example, a representative ABZI, compound 1, showed 59 ± 8% inhibition at 10 μM). Compound 1 (Fig. 1a) has an apparent inhibitory constant (IC₅₀^{APP}) of 14 ± 2 μM (Fig. 2b) and stabilizes the thermal unfolding of STING with a ΔT_m (difference between apo and ligand-bound STING in the inflection point at which 50% of protein is thermally unfolded) of 1.5 °C at doses of more than 16 μM (Fig. 1b), a hallmark of bona fide ligand binding for STING^{9,10}.

The cytoplasmic-facing CTD of STING is a homodimeric complex that interacts with cGAMP through a network of hydrogen bonds and water-mediated interactions within a large (1,400 nm³) binding pocket^{10,11}. To further characterize the binding of ABZI compounds to STING, we determined the structure of compound 1 in complex with the STING CTD at a resolution of 1.91 Å (Extended Data Table 2). Compound 1 binds in the cGAMP binding pocket with two bound molecules per STING dimer (Fig. 1c). Each molecule interacts with one STING subunit, spanning the entire side of the pocket without obvious contacts across the dimer interface. The 1-ethyl-3-methyl-1H-pyrazole-5-carboxamide moiety of compound 1 binds at the bottom of the pocket, with the methyl group projecting into a hydrophobic cleft made up of Leu159 and Thr267 of STING while the ethyl group makes no clear contacts with any portion of the protein (Fig. 1d). A key hydrogen bond is formed between the pyrazole nitrogen of compound 1 and the hydroxyl group of Ser162 of STING, while the carboxamide of compound 1 forms a hydrogen bond with Thr263. The terminal amide of compound 1 forms an H-bond network with Ser241, which is located at the base of an otherwise disordered and open lid domain of STING.

Inspection of the co-crystal structure revealed that the N-1 vectors of each ABZI molecule were close in space, projected across the STING dimer interface and lacked interactions with the protein. Therefore, we proposed that replacing the N1-hydroxyphenethyl moiety (N-1) with a linker between the two molecules to create a single dimeric ligand would afford a substantial increase in binding affinity (Fig. 2a). To test our linking hypothesis, we synthesized a diABZI (compound 2) and demonstrated that linked dimeric ABZIs (diABZI) enhanced binding by more than 1,000-fold with an IC₅₀^{APP} of 20 ± 0.8 nM.

This marked shift in potency is rooted in Jenck's principle, according to which the linked ABZIs reflect the sum of binding energies from the two unlinked ABZIs provided that 1) the binding orientation of unlinked ligands is maintained and 2) unfavourable interactions of the

¹Pattern Recognition Receptor DPU, GlaxoSmithKline, Collegeville, PA, USA. ²Immunology & Oncology DPU, GlaxoSmithKline, Collegeville, PA, USA. ³Cellzome, GlaxoSmithKline R&D, Heidelberg, Germany. ⁴Platform Technology & Science, GlaxoSmithKline, Collegeville, PA, USA. ⁵These authors contributed equally: Joshi M. Ramanjulu, G. Scott Pesiridis, Jingsong Yang. ⁶These authors jointly supervised this work: James Smothers, Axel Hoos, John Bertin. *e-mail: joshi.m.ramanjulu@gsk.com

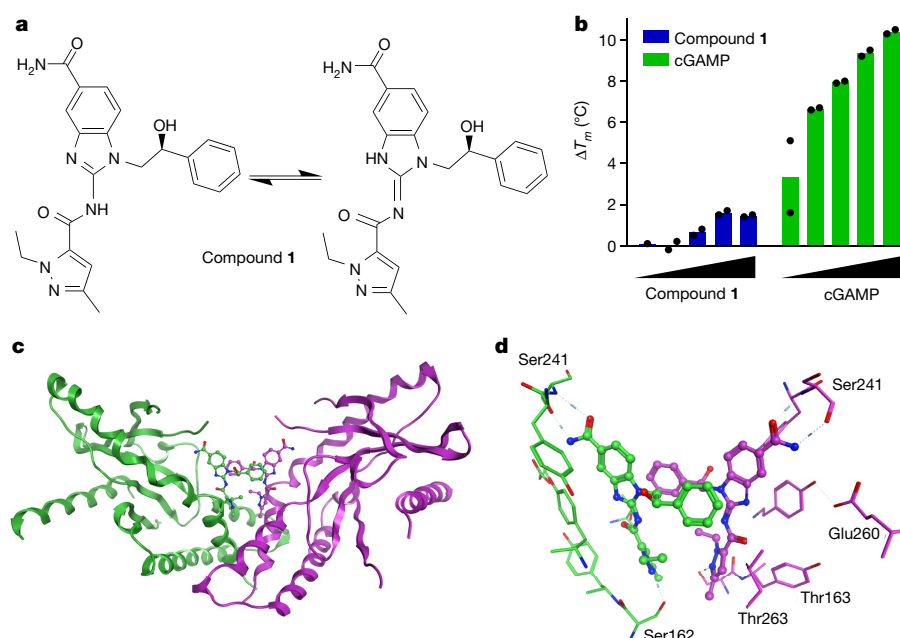


Fig. 1 | ABZI discovery and characterization. **a**, Tautomeric structure of ABZI compound 1. **b**, Compound 1 and cGAMP stabilize thermal unfolding of STING. Dose-dependent shift in T_m , inflection point at which 50% protein is thermally unfolded, induced by 1.8, 5.6, 16.6, and 50 μM

of compound 1 or cGAMP ($n = 2$). ΔT_m is difference between apo and ligand bound STING. **c**, Structure of the compound 1-STING complex; STING monomers are in green and purple. **d**, Close-up of the compound 1-STING complex depicting key interactions.

linker with the protein are avoided¹². A 2.4 Å resolution structure of the complex between compound 2 and STING confirmed that compound 2 maintains the same protein-ligand contacts that were observed with compound 1, and no interactions between the linker and the protein were observed (Extended Data Fig. 1). In addition to demonstrating a strong affinity to the CTD, compound 2 precipitated full-length STING from THP-1 cell lysates using a solid support immobilized with a derivative of compound 2 (compound 5) and detected by liquid chromatography with tandem mass spectrometry (LC-MS/MS). Competition with increasing concentrations of

compound 2 inhibited binding of full-length STING to the solid support with an apparent dissociation constant (K_d^{app}) of approximately 1.6 nM, a value consistent with the apparent K_d measured for the CTD alone (Fig. 1c, Extended Data Fig. 2). In summary, we devised a linking strategy that takes advantage of the symmetrical nature of STING and yielded a high-affinity ligand that interacts with human STING in a cGAMP-competitive manner.

Activation of STING through DNA sensing by the enzyme cGAS produces type I interferons and pro-inflammatory cytokines through direct activation by cGAMP^{11,13–15}. To determine the functional

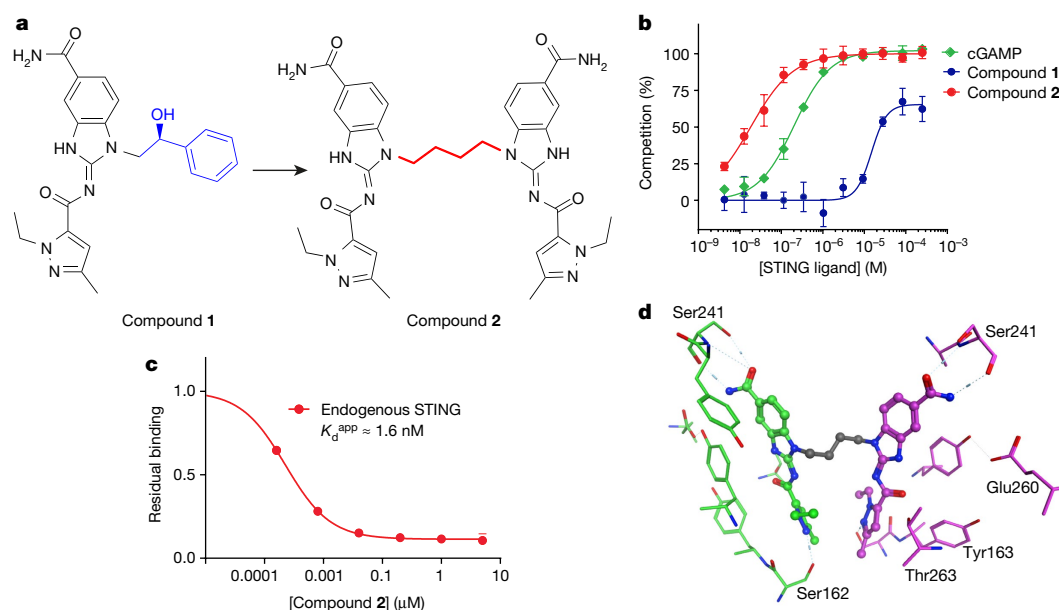


Fig. 2 | ABZI linking strategy and characterization of compound 2.

a, ABZI linking strategy illustrating replacement of hydroxyphenylethyl with 4-carbon butane linker to derive compound 2. **b**, Relative potency of compound 1 (ABZI) ($n = 3$), compound 2 (diABZI) ($n = 4$), and cGAMP ($n = 2$) measured by STING competition binding assay. Mean

response \pm s.d. **c**, Chemoproteomic analysis of THP-1 cell lysates using LC-MS/MS confirmed binding of endogenous full-length STING. Data represent mean response ($n = 2$). **d**, Close-up of diABZI compound 2 binding in the ligand-binding pocket of STING.

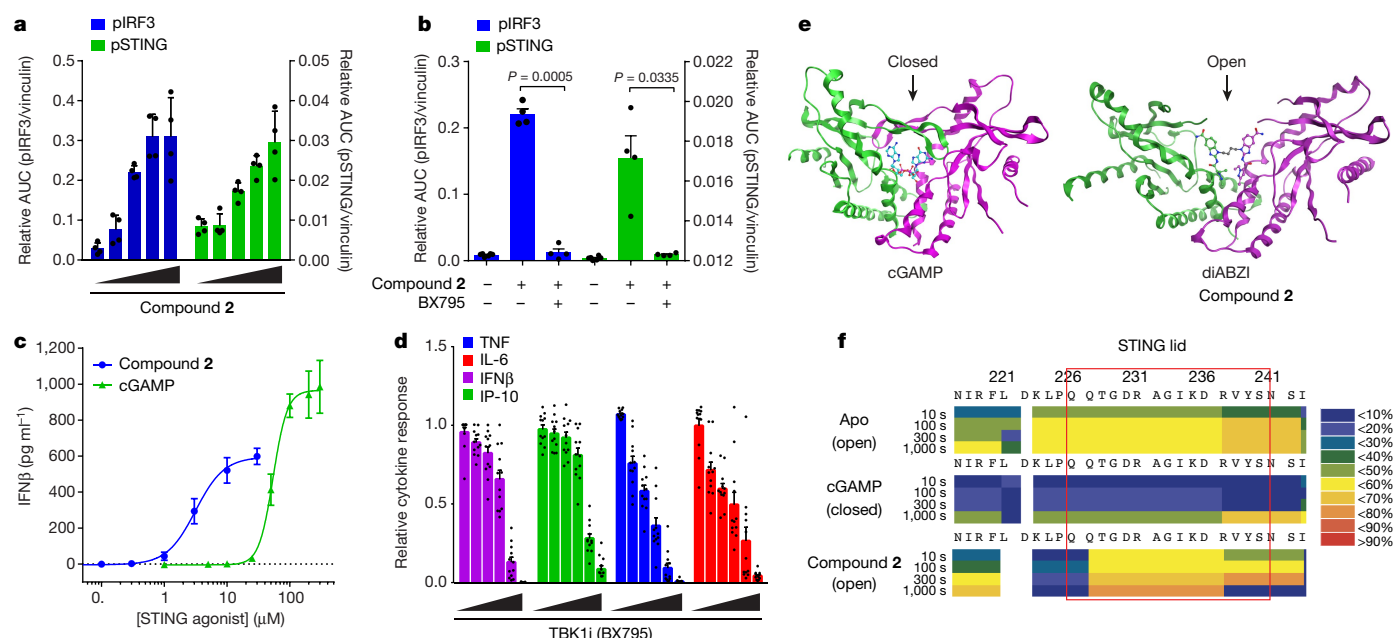


Fig. 3 | ABZI STING ligands are agonists. **a**, Dose-dependent phosphorylation of IRF3 and STING in human PBMCs following 2-h incubation with increasing concentrations of diABZI compound 2 (0.3, 1, 3, 10 and 30 μ M). **b**, The TBK1 inhibitor BX795 (10 μ M) inhibits phosphorylation of IRF3 and STING by compound 2 (3 μ M). Bars in **a**, **b** represent mean \pm s.e.m. IRF3 or STING phosphorylation compared to loading control protein vinculin from two donors measured in duplicate ($n = 4$). **b**, Paired t -test with threshold of $P < 0.05$. **c**, Human PBMCs treated with compound 2 or cGAMP demonstrate dose-dependent activation of STING with secretion of IFN β . Agonist model with fit line and error bars representing s.e.m. from replicate responses in two donors measured in duplicate ($n = 4$) for cGAMP or six donors in duplicate or triplicate ($n = 12$) for compound 2. **d**, Maximum activation of IFN β , IP-10,

TNF, and IL-6 following 4-h incubation with 3 μ M (EC_{50}) compound 2. Response was inhibited by dose titration of TBK1 inhibitor BX795 (0.03, 0.1, 0.3, 1 and 3 μ M). Data represent mean \pm s.e.m. from four donors in triplicate ($n = 12$). **e**, Comparison of cGAMP–STING (PDB accession code 4KSY)⁹ and diABZI compound 2–STING structures. Arrows depict closed and open conformations of cGAMP-bound and compound 2-bound states, respectively. **f**, Conformational state of apo-STING and cGAMP-bound or compound 2-bound conformations determined by hydrogen–deuterium exchange mass spectrometry. Heat map reflects rate of hydrogen incorporation into deuterium-saturated STING measured after 10, 100, 300, and 1,000 s. Red box highlights the ‘lid’ loop (aa 218–241); a closed or protected conformation (blue) versus a more solvent-exposed or open conformation (yellow).

consequence of binding, we incubated human peripheral blood mononuclear cells (PBMCs) with compound 2 and measured phosphorylation of STING, phosphorylation of IRF3 and the secretion of cytokines. Compound 2 caused dose-dependent phosphorylation of IRF3 and STING (Fig. 3a) that was inhibited by the TBK1 inhibitor BX795 (Fig. 3b). Similar to cGAMP, compound 2 induced dose-dependent secretion of IFN β with an EC_{50}^{app} of 3.1 ± 0.6 μ M. Compound 2 is therefore around 18-fold more potent than cGAMP, which has an EC_{50}^{app} of 53.9 ± 5 μ M (Fig. 3c). Both cGAMP and compound 2 appeared to be much less potent than demonstrated by their high binding affinity (Fig. 2b), probably owing to their poor ability to cross the cell membrane¹⁰. In addition to IFN β , compound 2 promotes production of interferon γ -induced protein 10 (IP-10), interleukin 6 (IL-6) and tumour necrosis factor (TNF, also known as TNF α) by a mechanism that is dependent on STING-mediated activation of TBK1 (Fig. 3d).

Structural studies of STING bound to different cyclic dinucleotides and the STING ligand DMXAA suggest that ligands that induce the closed conformation of STING result in its activation^{11,16,17}. However, unlike cGAMP and DMXAA, compound 2 efficiently activated STING function while maintaining an open STING conformation (Fig. 3e). To confirm that STING maintains an open conformation when bound to compound 2 in solution and that this is not a consequence of crystal packing, we investigated the conformation of the CTD of STING using hydrogen–deuterium (HD)-exchange mass spectrometry in the presence of cGAMP or compound 2 and in the unbound state. As expected, the lid region between Q227 and Y240 showed rapid HD exchange in the apo-STING conformation and transitioned to a highly protected environment with slower HD exchange in the presence of cGAMP (Fig. 3f). This is consistent with the transition from the open to closed conformations seen in the known crystal structures of STING bound

to cGAMP (Fig. 3e). By contrast, binding of compound 2 did not cause a shift in HD exchange in the lid region, confirming that STING maintains the open conformation when bound to compound 2 (Fig. 3f). This raises the possibility that activation of STING does not require the closed conformation.

Superimposed structures of unbound STING and STING bound to cGAMP or compound 2 show that the conformation of STING when bound to compound 2 is similar to the open apo-state conformation (Extended Data Fig. 3), whereas the structure when bound to compound 2 differs (Extended Data Fig. 4). No obvious differences were detected in the base of the binding pocket to explain changes in residue position or conformation that drive activation. Similarly, the binding of c-di-GMP to an open STING conformation can also lead to activation^{18,19}. Furthermore, gain-of-function mutations of STING are located outside the lid domain but have been reported to cause cGAMP-independent activation^{20,21}. These observations cast doubt on the idea that lid closure is critical for activation. While cGAMP requires lid domain interactions for high-affinity binding and induction of the closed conformation, activation of STING by ABZI-based agonists supports a model that does not require the closure of the lid. More research is needed to better understand how STING conformation modulates pathway activation.

With further lead optimization, we identified a representative compound from the diABZI series, compound 3, that has high binding affinity, improved potency in primary cells and similar functional activity across different human haplotypes and mouse STING (Fig. 4a). In human PBMCs, compound 3 induced dose-dependent activation of STING and secretion of IFN β with an EC_{50}^{app} of 130 nM. This is more than 400-fold more potent than cGAMP. Because several kinases, such as TBK1, IKK, AMPK, and ULK1, can modulate the STING pathway, we conducted a ³³P-radiolabelled kinase screening assay to measure

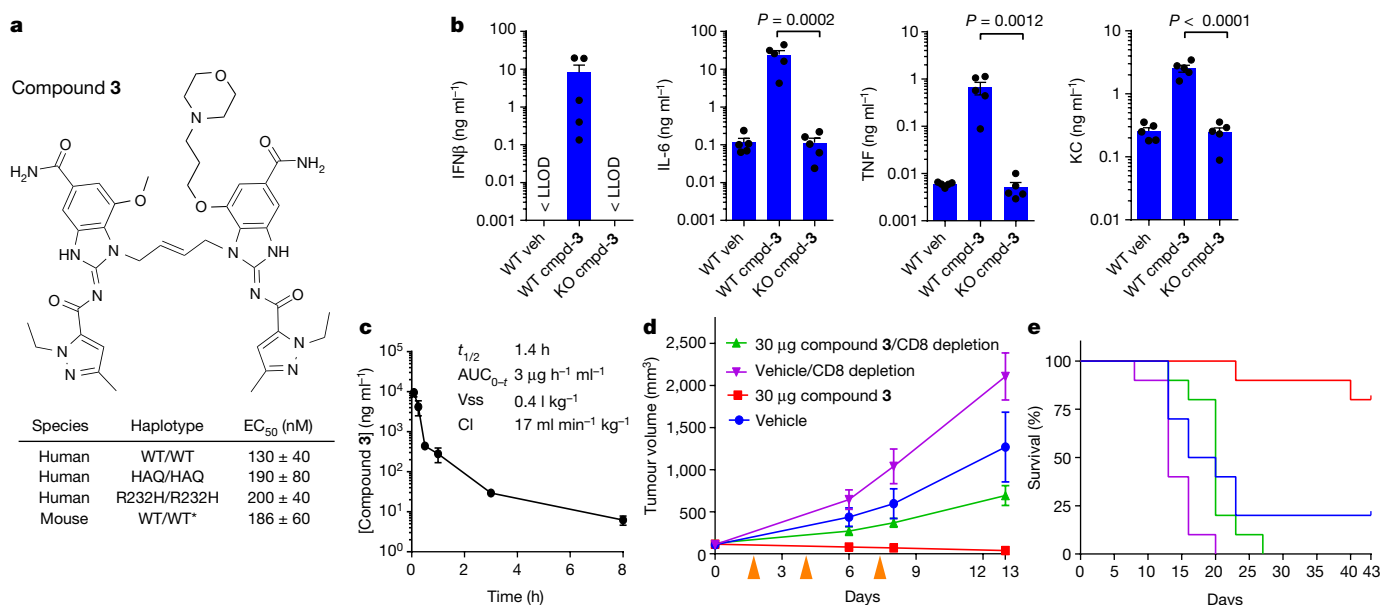


Fig. 4 | Systemic function of diABZI STING agonist 3. **a**, Structure of compound 3 (cmpd-3) and potency determined by treatment of PBMCs procured from humans with homozygous haplotypes of wild-type (WT; $n = 6$ from 3 donors), HAQ ($n = 2$ from 1 donor), R232H ($n = 2$ from 1 donor) and mouse (*combined WT/WT PBMCs). R71H-G230A-R293Q (HAQ) is the second most common human STING allele. Data represent mean EC₅₀ ± s.d. **b**, Compound 3 (2.5 mg kg⁻¹)-dependent secretion of IFNβ, TNF, IL-6, and KC/GROα in blood serum from wild-type or *Sting*^{-/-} mice. Statistical significance determined by one-way ANOVA with indicated P values. **c**, Pharmacokinetics of 3 mg kg⁻¹ compound 3 following intravenous bolus in BALB/c mice ($n = 5$). LLOD, lower limit of detection. Pharmacokinetic parameters reflecting half-life ($t_{1/2}$), area under the curve (AUC_{0-∞}), volume of distribution (V_{ss}) and clearance rate (Cl) are shown. Data represent mean ± s.d.m. of measured concentration of compound 3 ($n = 3$). **d**, Efficacy of intravenous injection of 1.5 mg kg⁻¹

of compound 3 (red) or vehicle (blue) in BALB/c mice bearing a single subcutaneous CT-26 colorectal tumour (~100 mm³). Mice with CD8⁺ cells depleted by pre-dose intraperitoneal injection of anti-CD8 300 μg antibody (15 mg kg⁻¹) (Extended Data Fig. 5), were similarly administered 30 μg (1.5 mg kg⁻¹) of compound 3 (green) or vehicle (purple). Mice were treated with three intravenous doses of compound 3 or vehicle on days 1, 4 and 8 (orange arrowheads) followed by measurements of tumour volume. Data represent mean tumour volume ± s.e.m. ($n = 10$ per group). Treatment group (30 μg) showed statistically significant reduction in tumour volume compared to all other groups using non-parametric ANOVA, $P < 0.001$. **e**, Kaplan–Myer survival plot of mice treated intravenously in **d** with 1.5 mg kg⁻¹ of compound 3 in the presence and absence of CD8⁺ cells. Kaplan–Myer log rank test demonstrated improved survival of 30 μg treatment group ($P < 0.001$ compared to all other groups).

selectivity. At a concentration of 1 μM, compound 3 demonstrated high selectivity against more than 350 kinases tested (Extended Data Table 3). This result, combined with additional cross-target selectivity screenings, indicates that diABZI STING agonists such as compound 3 are remarkably selective for STING.

Activation of STING elicits a type-I interferon response that propagates interferon receptor signalling in tumour-resident dendritic cells and leads to antitumour CD8⁺ T cell responses in vivo^{6,7}. To evaluate the in vivo activity of diABZIs, we treated wild-type C57Blk6 mice and *Sting*^{-/-} (also known as *Tmem173*^{-/-}) mice with compound 3 via subcutaneous injection. Compound 3 activated secretion of IFNβ, IL-6, TNF, and KC/GROα (also known as CXCL1) in wild-type but not *Sting*^{-/-} mice, confirming that compound 3 induces STING-dependent activation of type-I interferon and pro-inflammatory cytokines in vivo (Fig. 4b).

The murine STING agonist DMXAA elicits tumour regression upon intratumoral and intraperitoneal delivery²², suggesting that systemic delivery of a STING agonist engages anti-tumour mechanism(s) that drive tumour regression. Intravenous injection with high daily doses of cGAMP results in only modest in vivo efficacy²³. Intramuscular injection of cGAMP delayed tumour growth when used prophylactically, with injections started 4 days after tumour implantation²⁴. Therefore, there is a need to develop STING agonists that are active in humans and can induce potent efficacy when delivered systemically⁹.

To evaluate the potential therapeutic effects of systemically administered ABZI STING agonists, we tested the efficacy of intravenously delivered diABZI compound 3 in a syngeneic mouse model of colorectal tumours (CT-26) in BALB/c mice. We first established the pharmacokinetic profile of compound 3 in BALB/c mice following intravenous injection of 3 mg kg⁻¹ (Fig. 4c). Compound 3 exhibited systemic exposure with a half-life of 1.4 h and achieved systemic

concentrations greater than the half-maximal effective concentration (EC₅₀) for mouse STING (~200 ng ml⁻¹; Fig. 4a, c). Next, we tested an intermittent dosing paradigm in which 1.5 mg kg⁻¹ compound 3 was injected intravenously on days 1, 4, and 8 in mice with approximately 100 mm³ subcutaneous CT-26 tumours. Treatment with compound 3 resulted in significant tumour growth inhibition as measured by tumour volume AUC analysis ($P < 0.001$), and significantly improved survival ($P < 0.001$) with 8 out of 10 mice remaining tumour free at the end of the study on day 43 (Fig. 4d). To further dissect the mechanism of the anti-tumour activity of STING and to investigate the contribution of the immune system to the observed efficacy, we carried out a similar study in mice treated with an anti-mouse CD8 antibody to deplete CD8⁺ T cells (Extended Data Fig. 5). Depletion of CD8⁺ cells resulted in a significant decrease in the efficacy of intravenous injections of 1.5 mg kg⁻¹ of compound 3, in both tumour growth inhibition and survival benefit, with no tumour-free mice ($P < 0.001$; Fig. 4e). These data provide compelling evidence that activation of an adaptive immune response mediates the durable anti-tumour effect of compound 3 and causes complete tumour regression.

Current clinical trials of STING agonists are focused on intratumoral delivery. Aside from the technical challenge of intratumoral drug administration, the therapeutic potential of such STING agonists is limited to patients with accessible solid tumours. In addition, it is challenging to demonstrate a durable abscopal effect in patients with multiple heterogeneous, distal tumours. To overcome these challenges, we developed a small-molecule STING agonist, intravenous administration of which leads to an adaptive CD8⁺ T cell response in vivo. To our knowledge, diABZI compounds such as compound 3 represent the first potent, non-nucleotide STING agonists and have tremendous potential to improve treatment of cancer in humans.

Data availability

All data generated or analysed during this study are included in this published article (and its Supplementary Information files). Structure datasets generated during the current study are available in the PDB repository under accession numbers 6DXG and 6DXL. Additional data are available from the corresponding author on reasonable request.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0705-y>.

Received: 22 June 2018; Accepted: 15 October 2018;

Published online 7 November 2018.

- Corrales, L., McWhirter, S. M., Dubensky, T. W., Jr & Gajewski, T. F. The host STING pathway at the interface of cancer and immunity. *J. Clin. Invest.* **126**, 2404–2411 (2016).
- Li, T. & Chen, Z. J. The cGAS-cGAMP-STING pathway connects DNA damage to inflammation, senescence, and cancer. *J. Exp. Med.* **215**, 1287–1299 (2018).
- Mullard, A. Can innate immune system targets turn up the heat on 'cold' tumours? *Nat. Rev. Drug Discov.* **17**, 3–5 (2018).
- Woo, S. R. et al. STING-dependent cytosolic DNA sensing mediates innate immune recognition of immunogenic tumors. *Immunity* **41**, 830–842 (2014).
- Klarquist, J. et al. STING-mediated DNA sensing promotes antitumor and autoimmune responses to dying cells. *J. Immunol.* **193**, 6124–6134 (2014).
- Diamond, M. S. et al. Type I interferon is selectively required by dendritic cells for immune rejection of tumors. *J. Exp. Med.* **208**, 1989–2003 (2011).
- Fuertes, M. B. et al. Host type I IFN signals are required for antitumor CD8⁺ T cell responses through CD8alpha⁺ dendritic cells. *J. Exp. Med.* **208**, 2005–2016 (2011).
- Corrales, L. et al. Direct activation of STING in the tumor microenvironment leads to potent and systemic tumor regression and immunity. *Cell Reports* **11**, 1018–1030 (2015).
- Conlon, J. et al. Mouse, but not human STING, binds and signals in response to the vascular disrupting agent 5,6-dimethylxanthone-4-acetic acid. *J. Immunol.* **190**, 5216–5225 (2013).
- Ouyang, S. et al. Structural analysis of the STING adaptor protein reveals a hydrophobic dimer interface and mode of cyclic di-GMP binding. *Immunity* **36**, 1073–1086 (2012).
- Zhang, X. et al. Cyclic GMP-AMP containing mixed phosphodiester linkages is an endogenous high-affinity ligand for STING. *Mol. Cell* **51**, 226–235 (2013).
- Jencks, W. P. On the attribution and additivity of binding energies. *Proc. Natl Acad. Sci. USA* **78**, 4046–4050 (1981).
- Ishikawa, H. & Barber, G. N. STING is an endoplasmic reticulum adaptor that facilitates innate immune signalling. *Nature* **455**, 674–678 (2008).
- Ishikawa, H., Ma, Z. & Barber, G. N. STING regulates intracellular DNA-mediated, type I interferon-dependent innate immunity. *Nature* **461**, 788–792 (2009).
- Gao, P. et al. Cyclic [G(2',5')pA(3',5')p] is the metazoan second messenger produced by DNA-activated cyclic GMP-AMP synthase. *Cell* **153**, 1094–1107 (2013).
- Diner, E. J. et al. The innate immune DNA sensor cGAS produces a noncanonical cyclic dinucleotide that activates human STING. *Cell Reports* **3**, 1355–1361 (2013).
- Gao, P. et al. Structure-function analysis of STING activation by c[G(2',5')pA(3',5')p] and targeting by antiviral DMXAA. *Cell* **154**, 748–762 (2013).
- Kranzusch, P. J. et al. Ancient origin of cGAS-STING reveals mechanism of universal 2',3' cGAMP signaling. *Mol. Cell* **59**, 891–903 (2015).
- Shu, C., Yi, G., Watts, T., Kao, C. C. & Li, P. Structure of STING bound to cyclic di-GMP reveals the mechanism of cyclic dinucleotide recognition by the immune system. *Nat. Struct. Mol. Biol.* **19**, 722–724 (2012).
- Liu, Y. et al. Activated STING in a vascular and pulmonary syndrome. *N. Engl. J. Med.* **371**, 507–518 (2014).
- Melki, I. et al. Disease-associated mutations identify a novel region in human STING necessary for the control of type I interferon signaling. *J. Allergy Clin. Immunol.* **140**, 543–552 (2017).
- Jassar, A. S. et al. Activation of tumor-associated macrophages by the vascular disrupting agent 5,6-dimethylxanthone-4-acetic acid induces an effective CD8⁺ T-cell-mediated antitumor immune response in murine models of lung cancer and mesothelioma. *Cancer Res.* **65**, 11752–11761 (2005).
- Li, T. et al. Antitumor activity of cGAMP via stimulation of cGAS-cGAMP-STING-IRF3 mediated innate immune response. *Sci. Rep.* **6**, 19049 (2016).
- Wang, H. et al. cGAS is essential for the antitumor effect of immune checkpoint blockade. *Proc. Natl Acad. Sci. USA* **10.1073/pnas.1621363114** (2017).

Acknowledgements We thank B. Geddes for helpful suggestions and S. Romeril for discussions and comments.

Reviewer information Nature thanks B. Stockwell and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.M.R. conceived the dimer concept and designed compound **2**, and conceived the concept for compound **3** and synthetic chemistry for compound **4**. G.S.P. identified compound **1**. J.M.R., G.S.P. and J.Y. were co-leaders and oversaw the research program. J.M.R., G.S.P. and J.Y. wrote the manuscript with assistance from all other authors. N.C. performed HDX studies and determined X-ray structures with assistance from L.W. R.S. synthesized compounds **2** and **4**. S.-Y.Z., M.A., and C.B.H. conducted the in vivo efficacy study in CT-26 tumour-bearing mice. J.-L.T. conducted in vivo pharmacodynamics studies in wild-type and *Sting*^{−/−} mice. P.M. performed in vitro functional experiments in PBMCs. S.L., H.C.E., M.M., M.B., and G.B. designed, performed and analysed chemoproteomics experiments. M.K. and J.L.S. developed and assisted with the high-throughput screening assay. J.C. conducted PBMC assays from different haplotype donors. J.M., J.R., A.M., L.L., T.L.G., A.K.C., G.Y., and Y. Li contributed to design, optimization of synthetic route and preparation of compounds. N.N. carried out structure-based design analysis. A.I.W., V.K., and P.M. characterized agonist activity. K.N. purified STING protein. J.G. conducted thermal shift experiments. K.B. and M.A.R. designed and supervised pharmacokinetic studies. K.P.F. was co-leader during program initiation. P.J.G. supervised biology and provided advice. Y. Lian, K.J.D., and J.A. contributed to compound selection. R.W.M. contributed to chemistry strategy and provided advice. J.K. contributed to optimization of synthetic route and preparation of compounds. J.S., A.H. and J.B. jointly supervised the program.

Competing interests The authors declare no competing interests.

Additional information

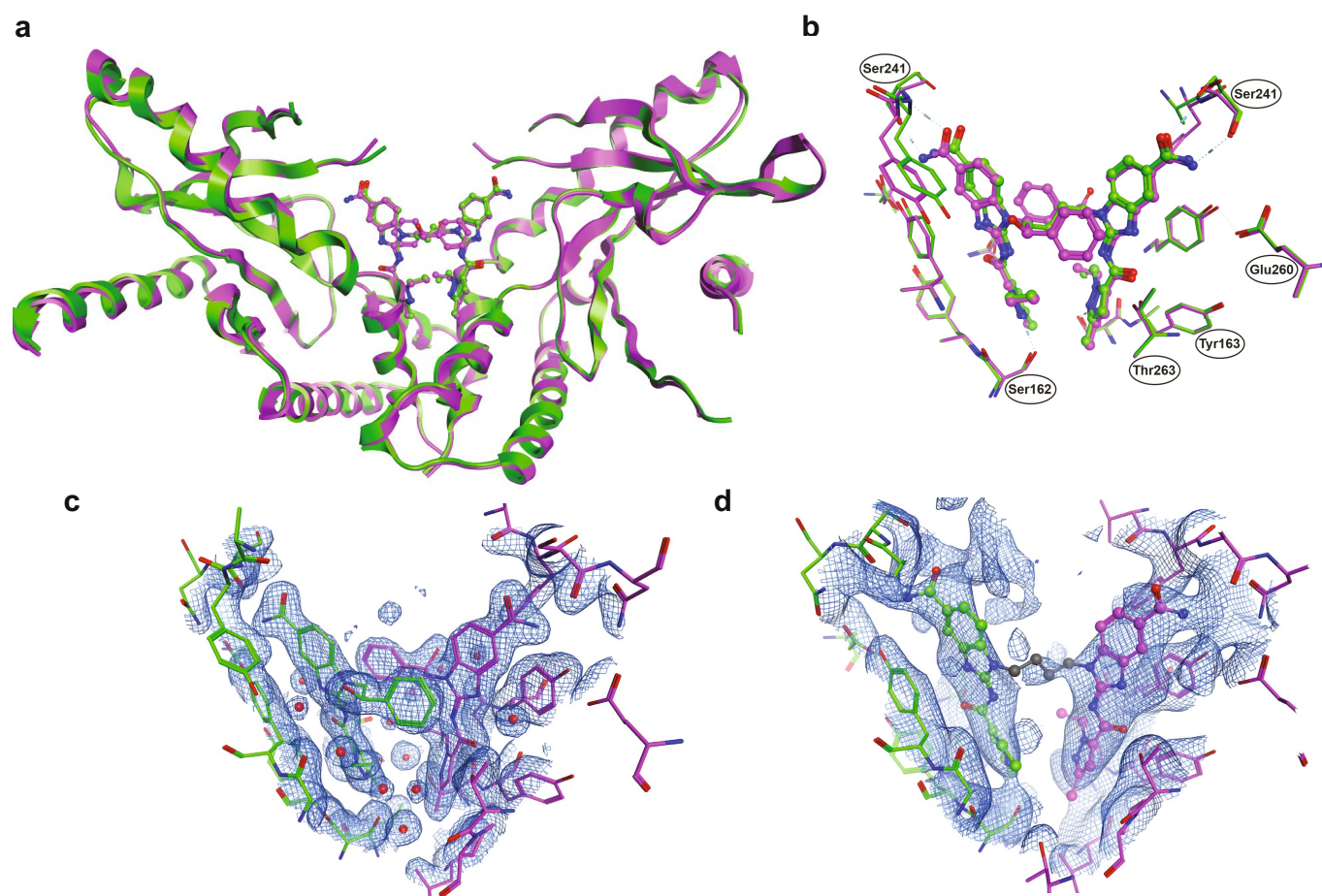
Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0705-y>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0705-y>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

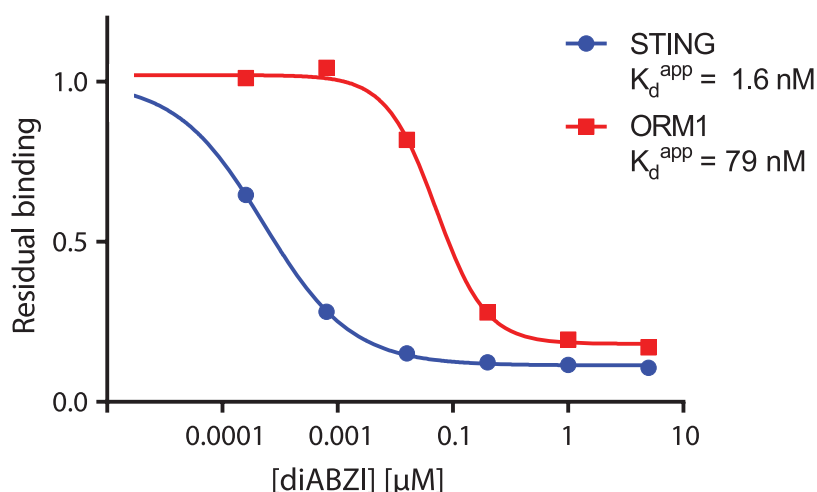
Correspondence and requests for materials should be addressed to J.M.R.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



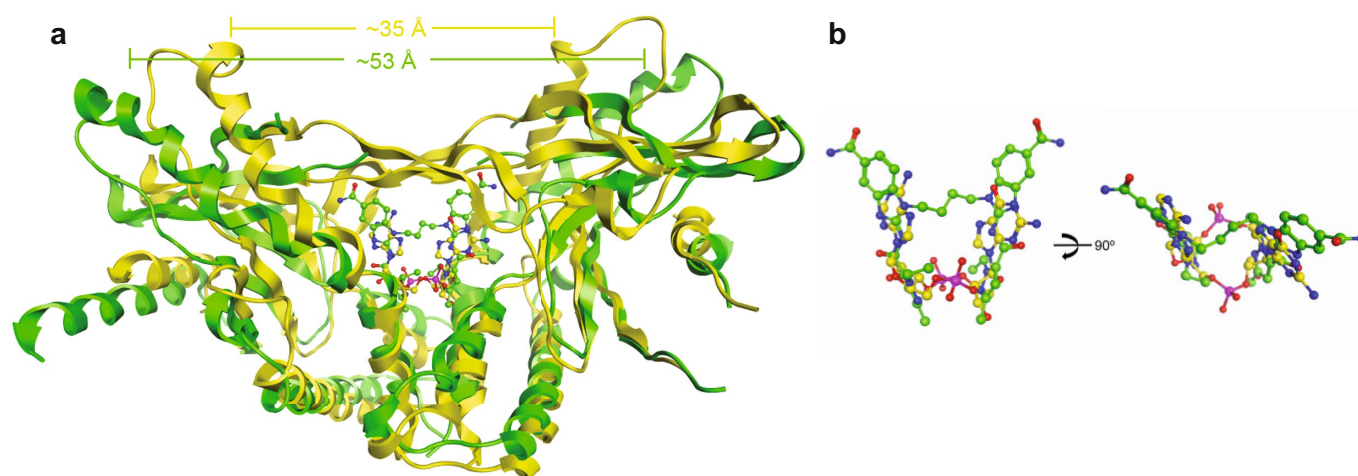
Extended Data Fig. 1 | Co-crystal structures and superimposition of compounds 1 and 2. **a**, Superposition of compound 1 (PDB code: 6DXG) and the diABZI compound 2 (PDB code: 6DXL) bound to human STING (aa 149–379). **b**, Intermolecular contacts in the complex of compounds 1

and 2 bound to human STING (aa 149–379). Magenta, compound 1; green, compound 2. Corresponding subunits of STING shown in same colour for compounds 1 and 2. **c**, Electron density (1.0 σ) of compound 1. **d**, Electron density of (0.5 σ) of compound 2.



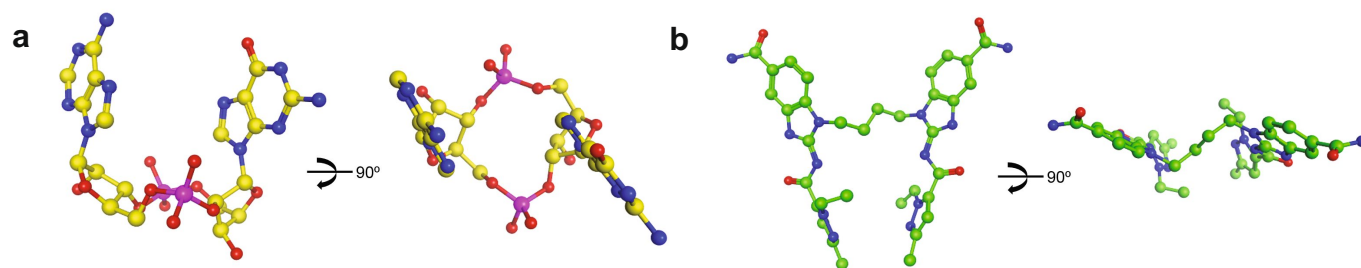
Extended Data Fig. 2 | Selectivity of compound 2 determined by affinity enrichment chemoproteomics. To identify any potential off-target liabilities early on, an affinity enrichment-based chemoproteomics strategy was applied to compound 2. Compound 5, an active analogue containing a primary amine functionality, was covalently immobilized on sepharose beads and was used to affinity-capture potential target proteins from a THP1 cell lysate. Pull-down experiments were performed in the absence of free compound 2 to delineate target proteins from background or in the presence of compound 2 over a range of concentrations. All proteins captured by the beads under the different conditions were eluted and subsequently quantified by isotope tagging of tryptic peptides followed by LC-MS/MS analysis to establish a competition-binding curve and determine a half-maximal inhibition (IC_{50}) value. The IC_{50} values obtained in these experiments represent a measure of target affinity, but are also affected by the affinity of the target for the bead-immobilized

ligand. The latter effect can be deduced by determining the depletion of the target proteins by the beads, such that apparent dissociation constants (K_d^{app}) can be determined, which are largely independent from the bead ligand (see Supplementary Methods for details). Notably, only two proteins were captured and competed in a dose-dependent manner within a 1,000-fold window, namely STING and orosomucoid1 (ORM1, alpha-1-acid glycoprotein 1 precursor). The mean K_d^{app} value for STING was determined as 1.6 nM, demonstrating high potency of compound 2 on the target protein not only in an artificial biochemical assay system using truncated protein but also against the full-length endogenous human protein. The mean K_d^{app} value of the only identified off-target protein, ORM1, was determined as 79 nM giving a comfortable selectivity window of approximately 40-fold. ORM1 is an acute phase reactant, an abundant plasma protein with known drug binding properties, and is known to be expressed in monocytes.

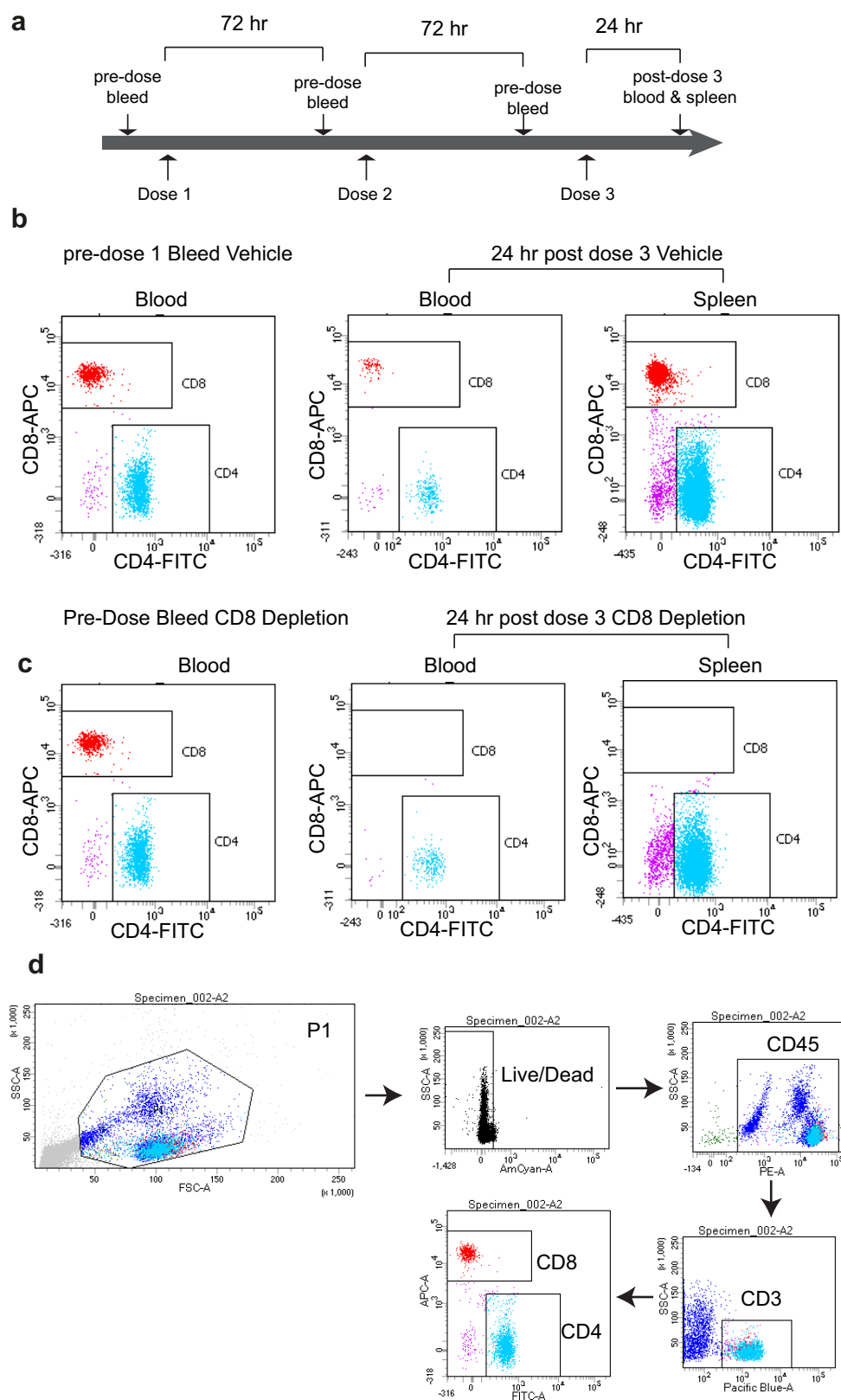


Extended Data Fig. 3 | Superimposition of co-crystal structures of cGAMP and compound 2. **a**, Superimposition of bound conformations of cGAMP (yellow) and diABZI compound 2 (green) bound to human

STING (aa 149–379). **b**, Superimposition of bound structures of cGAMP and diABZI compound 2.



Extended Data Fig. 4 | Bound conformations of cGAMP and compound 2. a, Conformations of cGAMP bound to human STING (aa 149–379). **b,** diABZI compound 2 bound to human STING (aa 149–379).



Extended Data Fig. 5 | Anti-CD8 depletion antibody validation by flow cytometry. **a**, Schematic of CD8T cell depletion scheme with timings consistent with efficacy studies. **b**, **c**, Flow cytometry quantification of CD4 and CD8 T cells from vehicle-treated (**b**) or anti-CD8 antibody (BioXcell: clone 2.43)-treated (**c**) BALB/c mice. Blood taken before dosing and after the third dose and spleen samples validate effective depletion of CD8⁺ T cells. Similar results observed 72 h after dose 1 and dose 2. **d**, Flow cytometry gating strategy. Flow cytometry staining and gating blood samples were collected via tail snip for pre-dose bleeds and via

cardiac puncture under isoflurane following the third dose. An equal volume of blood was added to flow staining buffer (PBS + 0.5% BSA), and samples were incubated in mouse Fc blocker. Spleen samples were processed to cell suspension, resuspended in flow staining buffer, and incubated with mouse Fc blocker. Samples were stained with live/dead aqua dye, CD45-PE, CD3-V421, and CD8-APC. Gating strategy reports the percentage positive population of live cells \rightarrow CD45⁺ \rightarrow CD3⁺. All samples were run on BD Canto II and analysed with FACSDiva software.

Extended Data Table 1 | Screening statistics and results for compound 1

| STING HTS Statistics | |
|--|------|
| Robust Mean (% I) | -0.1 |
| Robust Stdev (% I) | 7.35 |
| Primary Hit Cutoff (% I ^{3SD}) | 20.9 |
| Hit Rate (%) | 0.4 |
| Compound 1 HTS Data | |
| Primary HTS response (% I) | 67.7 |
| Hit Confirmation (% I) Replicate 1 | 54.9 |
| Hit Confirmation (% I) Replicate 2 | 53.8 |
| Average Response (%I) | 58.8 |
| Standard Deviation of Response (% I) | 7.70 |

Approximately 1.8×10^6 compounds from the GlaxoSmithKline (GSK) small-molecule screening collection were screened at a concentration of 10 μ M in 1,536-well plates using the ³H-cGAMP SPA assay. HTS, high-throughput screen.

Extended Data Table 2 | X-ray diffraction data collection and refinement statistics

| | STING-compound 1 complex | STING-compound 2 complex |
|--|-----------------------------|-----------------------------|
| Data collection | | |
| Space group | C222 ₁ | P 2 ₁ |
| Cell dimensions | | |
| <i>a</i> , <i>b</i> , <i>c</i> (Å) | 81.61, 92.22, 72.74 | 60.25, 73.14, 60.35 |
| α , β , γ (°) | 90.0, 90.0, 90.0 | 90.0, 96.19, 90.0 |
| Resolution (Å) | 1.91 (1.97 - 1.91) | 2.45 (2.54 - 2.45) * |
| <i>R</i> _{sym} or <i>R</i> _{merge} | 0.059 (0.690) | 0.080 (0.895) |
| <i>I</i> / σI | 20.92 (3.06) | 8.41 (1.62) |
| Completeness (%) | 97.36 (82.59) | 88.29 (53.26) |
| Redundancy | 7.8 (6.2) | 3.5 (3.2) |
| Refinement | | |
| Resolution (Å) | 1.91 | 2.45 |
| No. reflections | 21,239 | 17,065 |
| <i>R</i> _{work} / <i>R</i> _{free} | 0.194/0.203 | 0.247/0.248 |
| No. atoms | | |
| Protein | 1,404 | 2,732 |
| Ligand/ion | 32 | 52 |
| Water | 64 | 15 |
| <i>B</i> -factors | | |
| Protein | 54.5 | 81.1 |
| Ligand/ion | 46.7 | 112.4 |
| Water | 51.1 | 76.9 |
| R.m.s. deviations | | |
| Bond lengths (Å) | 0.008 | 0.005 |
| Bond angles (°) | 0.92 | 0.80 |

*Values in parentheses are for highest-resolution shell. The diffraction data for each dataset were collected from a single crystal.

Extended Data Table 3 | Kinome inhibition

| RBC Gene Name | Compound 3 (1 uM) | RBC Gene Name | Compound 3 (1 uM) | RBC Gene Name | Compound 3 (1 uM) | RBC Gene Name | Compound 3 (1 uM) |
|-------------------------|-------------------|---------------|-------------------|-----------------|-------------------|---------------|-------------------|
| ABL1 | -2.95 | DMPK | -3.18 | MAK | -3.71 | PKCTHETA | 8.725 |
| ABL2/ARG | 2.81 | DMPK2 | 2.8 | MAPKAPK2 | -2.15 | PKCZETA | -3.2 |
| ACK1 | 7.745 | DRAK1/STK17A | -5.145 | MAPKAPK3 | 3.005 | PKD2/PRKD2 | 1.8 |
| AKT1 | 1.695 | DYRK1/DYRK1A | -2.53 | MAPKAPK5/PRAK | -5.605 | PKG1A | -5.77 |
| AKT2 | 5.96 | DYRK1B | 5.71 | MARK1 | -5.775 | PKG1B | 2.105 |
| AKT3 | 2.325 | DYRK2 | -1.545 | MARK2/PAR-1Ba | -0.945 | PKG2/PRKG2 | -4.645 |
| ALK | -2 | DYRK3 | -11.65 | MARK3 | -7.95 | PKN1/PRK1 | -10.88 |
| ALK1/ACVRL1 | 0.975 | DYRK4 | -6.22 | MARK4 | 8.34 | PKN2/PRK2 | -2.455 |
| ALK2/ACVR1 | -6.975 | EGFR | 0.98 | MEK1 | -2.88 | PKN3/PRK3 | 1.705 |
| ALK3/BMPRI1A | -7.265 | EPHA1 | 15.735 | MEK2 | -8.955 | PLK1 | 1.5 |
| ALK4/ACVR1B | 6.87 | EPHA2 | -1.1 | MEK3 | -12.405 | PLK2 | 1.445 |
| ALK5/TGFBRI | 5.83 | EPHA3 | 7.98 | MEK5 | -1.425 | PLK3 | 5.905 |
| ALK6/BMPRI1B | -1.855 | EPHA4 | -3.925 | MEK11 | 8.485 | PLK4/SAK | -7.285 |
| ARAF | 7.615 | EPHA5 | 8.495 | MEKK2 | -15.475 | PRKX | 1.865 |
| ARMS/NUAK1 | 6.165 | EPHA6 | 5.775 | MEKK3 | -7.045 | PYK2/FAK2 | -2.125 |
| ASK1/MAP3K5 | -13.06 | EPHA7 | 2.425 | MEKK6 | 0.285 | RAF1 | -1.08 |
| AURORA A | -0.125 | EPHA8 | 1.035 | MELK | 69.51 | RET | -4.86 |
| AURORA B | -1.635 | EPHB1 | -3.49 | MINK/MINK1 | 8.465 | RIPK2 | 18.135 |
| AURORA C | -6.425 | EPHB2 | 0.215 | MKK4 | -8.385 | RIPK3 | 3.205 |
| AXL | -4.59 | EPHB3 | 2.615 | MKK6 | -12.27 | RIPK4 | 3.965 |
| BLK | 8.02 | EPHB4 | 0.925 | MKK7 | -4.045 | RIPK5 | 2.435 |
| BMPRI2 | -5.05 | ERBB2/HER2 | 0.475 | MLCK/MYLK | 1.56 | ROCK1 | 3.165 |
| BMX/ETK | 2.065 | ERBB4/HER4 | 13.295 | MLCK2/MYLK2 | -0.785 | ROCK2 | -3.475 |
| BRAF | 10.75 | ERK1 | 5.415 | MLK1/MAP3K9 | 2.79 | RON/MST1R | -6.885 |
| BRK | 10.98 | ERK2/MAPK1 | -3.035 | MLK2/MAP3K10 | 2.57 | ROS/ROS1 | 8.84 |
| BRSK1 | -1.15 | ERK5/MAPK7 | -0.735 | MLK3/MAP3K11 | 8.56 | RSK1 | -2.295 |
| BRSK2 | -1.51 | ERK7/MAPK15 | -10.13 | MLK4 | -2.625 | RSK2 | 2.155 |
| BTX | -1.31 | ERN1/IRE1 | -4.1 | MNK1 | -8.095 | RSK3 | -0.68 |
| C-KIT | 1.14 | ERN2/IRE2 | 15.56 | MNK2 | 2.745 | RSK4 | 1.005 |
| C-MER | 8.865 | FAK/PTK2 | -7.095 | MRCKa/CDC42BPA | -4.785 | SBK1 | 3.89 |
| C-MET | -0.625 | FER | -5.18 | MRCKb/CDC42BPB | -8.68 | SGK1 | 0.265 |
| C-SRC | -8.03 | FES/FPS | 2.535 | MSK1/RPS6KA5 | -4.54 | SGK2 | 3.35 |
| CAMK1A | 23.58 | FGFR1 | 3.545 | MSK2/RPS6KA4 | -0.845 | SGK3/SGKL | 4.835 |
| CAMK1B | 3.84 | FGFR2 | 7.06 | MSSK1-STK23 | 0.705 | SIK1 | -3.58 |
| CAMK1D | -2.675 | FGFR3 | -0.3 | MST1/STK4 | -3.49 | SIK2 | -3.61 |
| CAMK1G | 1.29 | FGFR4 | 3.55 | MST2/STK3 | -6.675 | SIK3 | -2.04 |
| CAMK2A | -3.625 | FGR | -1.905 | MST3/STK24 | 8.98 | SLK/STK2 | -2.23 |
| CAMK2B | 5.025 | FLT1-VEGFR1 | 1.32 | MST4 | 16.705 | SNARK/NUAK2 | 3.68 |
| CAMK2D | 2.53 | FLT3 | 2.685 | MUSK | -3.455 | SNRK | -0.015 |
| CAMK2G | -8.78 | FLT4-VEGFR3 | 1.88 | MYLK3 | -6.805 | SRMS | 5.405 |
| CAMK4 | -4.53 | FMS | -3.085 | MYLK4 | 3.525 | SRPK1 | -8.985 |
| CAMKK1 | -4.95 | FRK-PTK5 | 3.62 | MYO3A | -7.71 | SRPK2 | 0.475 |
| CAMKK2 | 13.54 | FYN | -6.55 | MYO3B | 2.455 | SSTK/TSSK6 | -2.37 |
| CDC7/DBF4 | -3.395 | GCK/MAP4K2 | 8.355 | NEK1 | -0.715 | STK16 | -7.255 |
| CDK1/cyclin A | 0.63 | GLK/MAP4K3 | -0.74 | NEK11 | 4.665 | STK21/CIT | 0.105 |
| CDK1/cyclin B | 1.03 | GRK1 | 9.27 | NEK2 | 4.58 | STK22D/TSSK1 | -6.42 |
| CDK1/cyclin E | -1.11 | GRK2 | -6.385 | NEK3 | 2.34 | STK25/YSK1 | 11.52 |
| CDK14/cyclin Y (PFTK1) | 5.905 | GRK3 | -3.28 | NEK4 | 1.865 | STK32B/YANK2 | 4.165 |
| CDK16/cyclin Y (PCTAIR) | -3.73 | GRK4 | 9.195 | NEK5 | -11.455 | STK32C/YANK3 | -11.465 |
| CDK17/cyclin Y (PCTK2) | -3.32 | GRK5 | -5 | NEK6 | 9.665 | STK33 | -14.15 |
| CDK18/cyclin Y (PCTK3) | -0.935 | GRK6 | -14.005 | NEK7 | -19.95 | STK38-NDIR1 | 1.155 |
| CDK19/cyclin C | -1.955 | GRK7 | 10.71 | NEK8 | -11.715 | STK38L-NDIR2 | 0.365 |
| CDK2/cyclin A | -8.61 | GSK3A | -4.645 | NEK9 | -1.15 | STK39/STLK3 | -15.6 |
| CDK2/Cyclin A1 | 1.11 | GSK3B | -5.535 | NIM1 | -7.83 | SYK | -6.26 |
| CDK2/cyclin E | 2.295 | HASPIN | 3.905 | NLK | 13.95 | TAK1 | 0.845 |
| CDK2/cyclin O | -1.67 | HCK | -1.6 | OSR1/OXSR1 | -2.925 | TAOK1 | 1.075 |
| CDK3/cyclin E | -8.23 | HGK/MAP4K4 | 3.38 | P38a/MAPK14 | 0.535 | TAOK2-TAO1 | 1.215 |
| CDK4-cyclin D1 | -9.975 | HIPK1 | 4.865 | P38b/MAPK11 | -5.585 | TAOK3-JIK | 11.045 |
| CDK4-cyclin D3 | -4.725 | HIPK2 | 4.905 | P38d/MAPK13 | 10.275 | TBK1 | -9.49 |
| CDK5/p25 | 1.225 | HIPK3 | -7.735 | P38G | 5.785 | TEC | -0.43 |
| CDK5/p35 | -2.885 | HIPK4 | 1.975 | p70S6K/RPS6KB1 | -8.7 | TESK1 | 2.925 |
| CDK6-cyclin D1 | -13.815 | HPK1/MAP4K1 | -3.7 | p70S6Kb-RPS6KB2 | -4.145 | TESK2 | -4.65 |
| CDK6-cyclin D3 | 2.715 | IGF-1R | -2.64 | PAK1 | 7.545 | TGFBRI2 | -10.36 |
| CDK7/cyclin H | -3.425 | IKKa/CHUK | 13.885 | PAK2 | -1.355 | TIE2/TEK | 8.355 |
| CDK9-cyclin K | 4.465 | IKKb/IKKB | 0.595 | PAK3 | 7.365 | TLK1 | -5.255 |
| CDK9/cyclin T1 | -9.565 | IKKe/IKBKE | -1.81 | PAK4 | 1.505 | TLK2 | -4.035 |
| CDK9/cyclin T2 | 3.19 | IR | -10.495 | PAK5 | 8.86 | TNIK | -6.19 |
| CHK1 | 4.465 | IRAK1 | -18.39 | PAK6 | -0.7 | TNK1 | 4.45 |
| CHK2 | 0.065 | IRAK4 | -9.91 | PASK | 1.395 | TRKA/NTRK1 | 5.815 |
| CK1A1 | 4.98 | IRR/INSRR | 14.24 | PBK/TOPIK | 1.17 | TRKB | -6.62 |
| CK1A1L | -7.49 | ITK | -3.26 | PDGFRA | 1.755 | TRKC | -1.725 |
| CK1D | -8.725 | JAK1 | 4.14 | PDGFRB | 3.725 | TSSK2 | -1.355 |
| CK1EPSILON | 7.235 | JAK2 | -1.585 | PDK1/PDPK1 | 10.72 | TSSK3/STK22C | -10.44 |
| CK1G1 | -5.355 | JAK3 | -3.355 | PEAK1 | -2.71 | TBKI | 2.575 |
| CK1G2 | 0.875 | JNK1 | -3.955 | PHKG1 | 5.975 | TBKI2 | 4.305 |
| CK1G3 | 2.29 | JNK2 | 5.82 | PHKG2 | -0.625 | TKX | -2.76 |
| CK2A | 6.15 | JNK3 | 8.865 | PIM1 | -1.785 | TYK1/LTK | -0.41 |
| CK2A2 | -5.185 | KDR/VEGFR2 | 5.64 | PIM2 | 3.99 | TYK2 | 12.05 |
| CLK1 | 9.805 | KHS/MAP4K5 | 0.105 | PIM3 | -3.3 | TYRO3-SKY | 9.385 |
| CLK2 | 11.775 | KSR1 | 3.905 | PKA | -0.7 | ULK1 | -5 |
| CLK3 | 16.61 | KSR2 | -1.01 | PKACB | 9.475 | ULK2 | -1.47 |
| CLK4 | 0.96 | LATS1 | 5.38 | PKACG | -4.51 | ULK3 | 4.135 |
| COT1/MAP3K8 | 5.325 | LATS2 | -2.02 | PKCA | -5.235 | VRK1 | -6.725 |
| CSK | -1.615 | LCK | 6.54 | PKCB1 | -2.895 | VRK2 | 0.125 |
| CTK-MATK | -2.155 | LCK2/ICK | 5.225 | PKCB2 | -3.565 | WEE1 | 6.58 |
| DAPK1 | -3.24 | LIMK1 | -9.38 | PKCD | -1.785 | WNK1 | 2.585 |
| DAPK2 | -3.15 | LIMK2 | -2.035 | PKCEPSILON | 2.22 | WNK2 | 12.575 |
| DCAMKL1 | 3.78 | LKB1 | -3.265 | PKCETA | 5.335 | WNK3 | -10.575 |
| DCAMKL2 | 0.095 | LOK/STK10 | -5.3 | PKCG | -13.595 | YES/YES1 | -10.14 |
| DDR1 | 7.265 | LRRK2 | -9.725 | PKCIOTA | 8.475 | YSK4/MAP3K19 | -3.575 |
| DDR2 | -4.19 | LYN | -2.785 | PKCmu/PRKD1 | -14.565 | ZAK/MLTK | 3.45 |
| DLK/MAP3K12 | 5.725 | LYN B | 5.035 | PKCnu/PRKD3 | -3.425 | ZAP70 | 4.14 |
| | | | | | | ZIPK/DAPK3 | 11.435 |

Per cent response with 1 μ M diABZI compound 3.

Controlling orthogonal ribosome subunit interactions enables evolution of new function

Wolfgang H. Schmied^{1,4}, Zakir Tnimov^{1,4}, Chayasith Uttamapinant^{1,2,4}, Christopher D. Rae¹, Stephen D. Fried^{1,3} & Jason W. Chin^{1*}

Orthogonal ribosomes are unnatural ribosomes that are directed towards orthogonal messenger RNAs in *Escherichia coli*, through an altered version of the 16S ribosomal RNA of the small subunit¹. Directed evolution of orthogonal ribosomes has provided access to new ribosomal function, and the evolved orthogonal ribosomes have enabled the encoding of multiple non-canonical amino acids into proteins^{2–4}. The original orthogonal ribosomes shared the pool of 23S ribosomal RNAs, contained in the large subunit, with endogenous ribosomes. Selectively directing a new 23S rRNA to an orthogonal mRNA, by controlling the association between the orthogonal 16S rRNAs and 23S rRNAs, would enable the evolution of new function in the large subunit. Previous work covalently linked orthogonal 16S rRNA and a circularly permuted 23S rRNA to create orthogonal ribosomes with low activity^{5,6}; however, the linked subunits in these ribosomes do not associate specifically with each other, and mediate translation by associating with endogenous subunits. Here we discover engineered orthogonal ‘stapled’ ribosomes (with subunits linked through an optimized RNA staple) with

activities comparable to that of the parent orthogonal ribosome; they minimize association with endogenous subunits and mediate translation of orthogonal mRNAs through the association of stapled subunits. We evolve cells with genomically encoded stapled ribosomes as the sole ribosomes, which support cellular growth at similar rates to natural ribosomes. Moreover, we visualize the engineered stapled ribosome structure by cryo-electron microscopy at 3.0 Å, revealing how the staple links the subunits and controls their association. We demonstrate the utility of controlling subunit association by evolving orthogonal stapled ribosomes which efficiently polymerize a sequence of monomers that the natural ribosome is intrinsically unable to translate. Our work provides a foundation for evolving the rRNA of the entire orthogonal ribosome for the encoded cellular synthesis of non-canonical biological polymers⁷.

The ribosome is a 2.5-megadalton molecular machine, composed of two subunits, that synthesizes proteins using mRNA templates⁸. Ribosomal translation represents the ultimate paradigm for the encoded, high-fidelity synthesis of long polymers of defined sequence

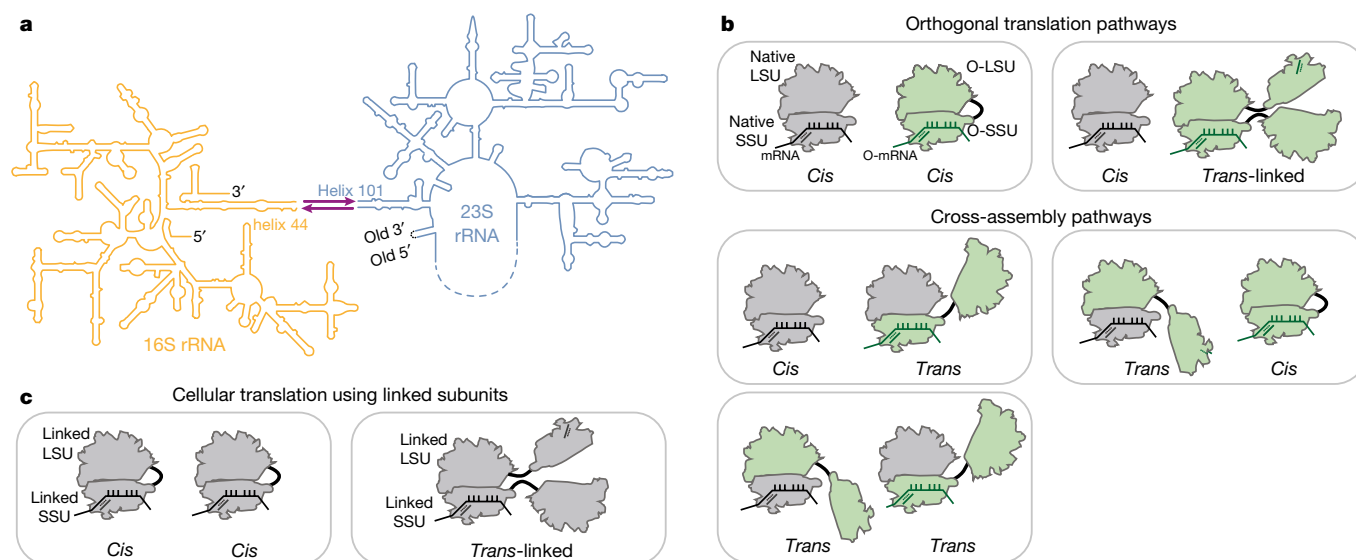


Fig. 1 | Ribosome stapling and potential interactions of linked subunits in vivo. **a**, Secondary structure of RNA from ribosomes with linked subunits. A single rRNA transcript is generated by inserting a circularly permuted 23S rRNA (blue) into a split 16S rRNA (yellow), with the 16S and 23S rRNAs linked together by an RNA linker (purple). In stapled ribosomes the linker is a staple derived from the J5–J5a region of the *Tetrahymena* group I intron. The original O-stapled ribosome, referred to here as O-d0d0, directly links h44 and H101 through the staple. **b**, Potential interactions of linked ribosomal subunits in vivo. Cells (white boxes) containing O-ribosomes with linked subunits as well as endogenous

(native) subunits may associate to create orthogonal translation pathways or cross-assembly pathways. An orthogonal translation complex is created by directing the association of an orthogonal small ribosome subunit (O-SSU) with its linked large subunit (O-LSU) in *cis*, or by forming *trans*-linked complexes. Linked ribosome subunits may interact with native ribosome subunits in *trans* if the linker is insufficient to direct assembly in *cis*. High concentrations of native ribosome subunits in the cytoplasm under physiological conditions counteract the effects of tethering and may lead to cross-assembly. **c**, In *E. coli* strains containing solely ribosomes with covalently linked subunits, *cis*- or *trans*-linked complexes may form.

¹Medical Research Council Laboratory of Molecular Biology, Cambridge, UK. ²Present address: School of Biomolecular Science and Engineering, Vidyasirimedhi Institute of Science and Technology (VISTEC), Rayong, Thailand. ³Present address: Department of Chemistry, Johns Hopkins University, Baltimore, MD, USA. ⁴These authors contributed equally: Wolfgang H. Schmied, Zakir Tnimov, Chayasith Uttamapinant. *e-mail: chin@mrc-lmb.cam.ac.uk

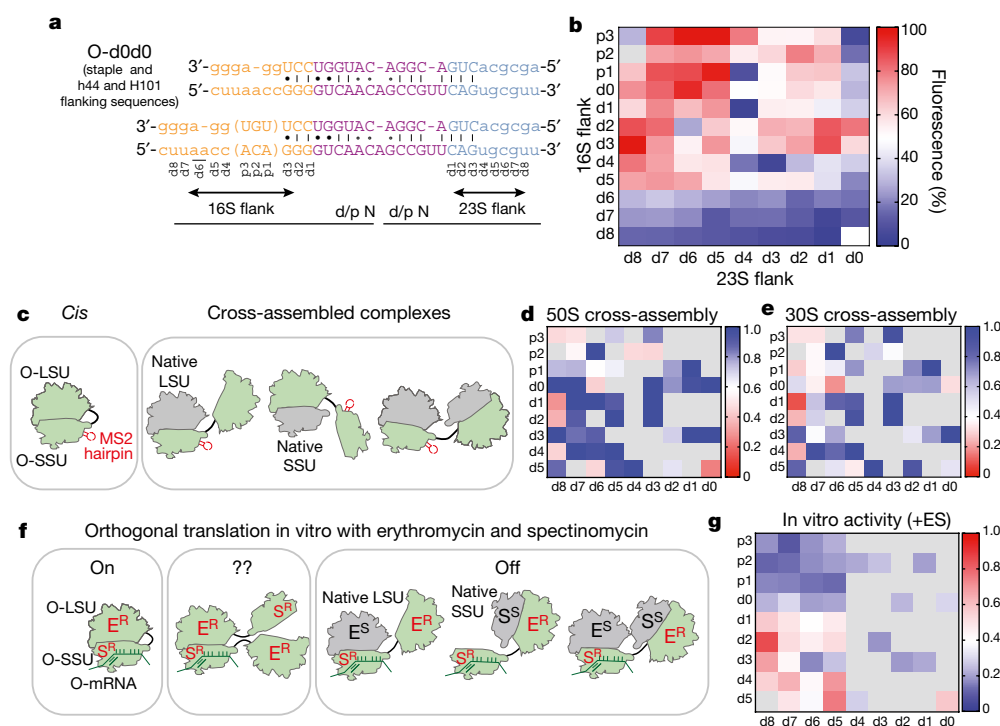


Fig. 2 | Maximizing activity and minimizing cross-assembly in engineered O-stapled ribosomes through systematic variation of the intersubunit linker. **a**, Top two rows, the intersubunit staple sequence (uppercase letters)—composed of a hinge (purple) and native helical flanking residues (yellow and blue)—used in O-d0d0. rRNA-derived sequences are in lowercase letters. Bottom, variants are denoted by the number (N) of base pairs that have been deleted (d) or inserted (p, for ‘plus’) from the 16S side, followed by the number of base pairs deleted from the 23S side. **b**, Heat map showing O-stapled ribosome activity (as analysed by the synthesis of GFP) in vivo. The resulting GFP fluorescence is shown as a percentage of that produced from an orthogonal ribosome with non-linked subunits. Data are thresholded at 100. See Methods for statistics and Extended Data Fig. 1b for full data. **c**, Potential complexes between small subunits (SSUs) and large subunits (LSUs) following affinity purification of O-stapled ribosomes (green) with an MS2 stem loop. Native

subunits are in grey. **d**, **e**, Heat maps showing 50S (LSU; **d**) and 30S (SSU; **e**) cross-assembly coefficients. Variants in grey were not tested. The heat map is thresholded at 1. Data are means of $n = 2$ biological replicates. See Extended Data Fig. 3a, b, f, In vitro translation of the orthogonal message in the presence of the antibiotics erythromycin and spectinomycin (which selectively inhibit native subunits) is denoted ‘on’, ‘off’ or unknown (‘??’) for each complex. E^R and S^R denote erythromycin and spectinomycin resistance; E^S and S^S denote erythromycin and spectinomycin sensitivity. **g**, Heat map showing the efficiency of translating T7-O-GFP (a construct containing a T7 promoter upstream of an orthogonal ribosome-binding site and an sfGFP gene) in S30 extracts of *E. coli*; the extracts contained the indicated O-stapled ribosome and native ribosomes. We added 10 μ M spectinomycin and 50 μ M erythromycin (ES) to the extract to inhibit native subunits. The heat map shows the mean fluorescence. For values of n and errors, see Methods and Extended Data Fig. 6.

and composition⁷. However, adapting the cellular ribosome to enable the encoded synthesis of non-canonical biopolymers is an outstanding challenge⁷. The small subunit of the ribosome, containing 16S rRNA, binds to mRNAs and decodes codon–anticodon interactions between the mRNA and transfer RNAs, whereas the large subunit, containing 23S rRNA, facilitates peptide-bond formation and co-translational protein folding; both subunits bind translation factors and coordinate translation⁸. Ribosomes are essential and many mutations in the ribosome are dominant-negative or lethal⁹.

In previous work, we created orthogonal (O-)ribosome–mRNA pairs in *E. coli*¹. In these pairs, an orthogonal message (O-mRNA, containing an O-ribosome-binding site) is selectively translated by the O-ribosome (containing O-16S rRNA with mutations in the anti-Shine–Dalgarno (ASD) sequence); this O-mRNA cannot, however, be translated by endogenous ribosomes¹. The O-ribosome is non-essential, and has been further evolved to enable efficient incorporation of multiple distinct non-canonical amino acids into polypeptides^{3,4}.

The orthogonal and wild-type ribosomes share a common pool of large subunits. An important goal is the creation of orthogonal ribosomes in which both a new 23S rRNA and the O-16S rRNA are directed towards an orthogonal message⁷. Such orthogonal ribosomes would maximize the contributions of the new 23S rRNA to translation of an O-mRNA, insulate the effects of otherwise deleterious mutations in the new 23S rRNA from cellular translation, and therefore enable the evolution of new function in the large subunit.

Ribosomal subunits can be covalently linked by joining helix 44 of the 16S rRNA to Helix 101 of a circularly permuted 23S rRNA (Fig. 1)^{5,6,10} through either a flexible tether, the $A_{8/9}$ tether, or the hinge from the J5–J5a region from the *Tetrahymena* group I self-splicing intron¹¹; the latter strategy created the parental orthogonal ‘stapled’ ribosome (herein called O-d0d0). These O-ribosomes with linked subunits maintain only 30% of the activity of the parental O-ribosome (Fig. 2 and Extended Data Fig. 1). Moreover, the exceptionally high concentration of ribosomes in cells (see Supplementary Information) suggests that subunit tethering—without additional features that favour association between the linked subunits and/or restrict the association of linked subunits with endogenous subunits—is unlikely to specify the association of linked subunits within an orthogonal ribosome (in *cis*) over association of the linked subunits with endogenous subunits (in *trans*) (Fig. 1b). Gain-of-function mutations in the 23S rRNA portion of an O-ribosome with tethered subunits led to a measurable phenotype when using an orthogonal message⁶. However, statistical partitioning of large subunits between orthogonal and endogenous small subunits can confer gain-of-function phenotypes without specificity in subunit association (Extended Data Fig. 1), and stronger gain-of-function phenotypes have been described with endogenous ribosomes^{12,13}. Orthogonal translation by linked subunits occurs in cells that also contain endogenous ribosomes; here, a different set of subunit associations may occur from those that may occur in cells that contain only ribosomes with covalently linked subunits that are directed to endogenous mRNAs (Fig. 1b, c). Therefore, in contrast to common assumptions, experiments that use

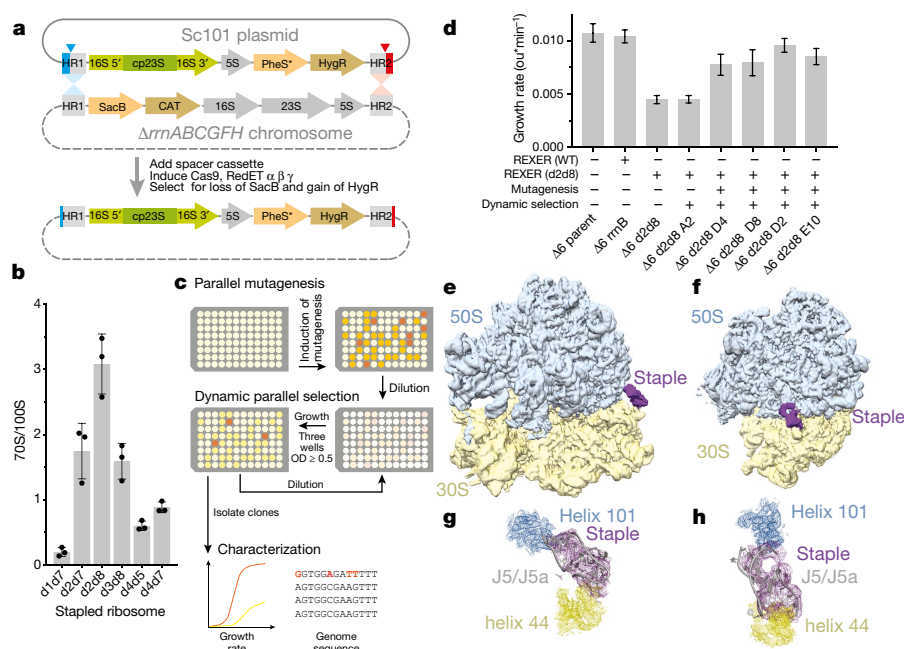


Fig. 3 | Genomically encoding stapled ribosomes as the sole cellular ribosomes, and subsequent strain evolution, generates fast-growing *E. coli* for d2d8 purification and structure determination. **a**, Schematic showing how ribo-REXER is used for the genomic replacement of ribosomal RNA operons, applied here to generate $\Delta 6$ d2d8. The SC101 plasmid contains the rRNA for the stapled ribosome; this rRNA is used to replace the single rRNA in the chromosome of a $\Delta rrnABCGFH$ strain of *E. coli*. *SacB*, *Bacillus subtilis* levansucrase gene; *CAT*, chloramphenicol acetyltransferase gene; cp, circularly permuted; HR, homologous regions; *HygR*, hygromycin-resistance gene; *PheS**, T251A A294G mutant of *E. coli* phenylalanyl-tRNA synthetase. Red and blue arrows indicate sites of Cas9-mediated cleavage. **b**, Ratio of ribosome monosomes (70S) to ribosome dimers (100S) from sucrose gradient analyses of different stapled ribosomes isolated from cells under associating conditions. Shown are individual data points (black dots), means (grey bars) and standard deviations from three biological replicates. **c**, Evolution of $\Delta 6$ d2d8 by

ribosomes with linked subunits, directed to endogenous messages, as the sole ribosomes in the cell (Fig. 1c) (which currently require muti-copy plasmids for rRNA expression and lead to retarded growth⁶) do not address the key question of whether orthogonal ribosomes with covalently linked subunits associate with the free subunits of endogenous ribosomes when both types of ribosome are present in the cell (Fig. 1b). Thus, there is no compelling evidence that the linked subunits within O-ribosomes reported to date specifically associate with each other to mediate translation⁷; nor is there any evidence that O-ribosomes can be altered to access new large-subunit functions that have not been accessed in natural ribosomes.

Here we identify engineered O-stapled ribosomes that minimize association with endogenous ribosomal subunits and maximize their activity through stapled-subunit association. In an evolved strain of *E. coli*, the engineered stapled ribosome supports robust cell growth as the sole, genomically encoded, ribosome, with growth rates comparable to those conferred by wild-type ribosomes. Cryo-electron microscopy (cryo-EM) reveals how the staple covalently links ribosome subunits to control their association, and we evolve engineered O-stapled ribosomes with new intrinsic polymerization function.

We envisioned that both the activity of the O-stapled ribosome directed to an orthogonal message, and the contributions to that activity from orthogonal translation pathways versus cross-assembly pathways (Fig. 1b, c), may vary as a function of the length of the helices in each subunit that link to the J5–J5a hinge (Fig. 2a). We created a matrix of 107 O-rDNAs (Fig. 2b, Extended Data Fig. 1b and Supplementary Data 1) that systematically combines deletions or insertions in helix 44 with deletions in Helix 101 in the O-stapled ribosome; these alterations

are expected to both translocate and rotate one subunit with respect to another. We named the linker variants by the number of base pairs that were deleted (d) or added (p, for ‘plus’) to helix 44 followed by the number of base pairs that have been deleted from Helix 101 with respect to O-d0d0 (Fig. 2a). We found that removing up to five nucleotides of helix 44 in O-16S rRNA leads to more active ribosomes, and that shortening Helix 101 of the 23S rRNA in the O-stapled ribosome is commonly associated with a gradual increase in activity. Indeed, some of the new O-stapled ribosomes are substantially more active than the first-generation O-ribosomes with linked subunits, and approach the activity of the non-stapled O-ribosome.

We next investigated the cross-assembly of O-stapled ribosomes and endogenous subunits (Fig. 1b) by affinity-purifying O-stapled ribosomes tagged with the RNA stem loop from the MS2 bacteriophage (refs^{14,15}) from cells and measuring the co-purification of endogenous subunits (Fig. 2c–e, Extended Data Fig. 2 and Supplementary Data 2). We defined the molar ratios of 50S and 30S rRNA to the MS2-tagged stapled O-rRNA from the purification as the 30S and 50S cross-assembly coefficients, respectively. We found that different O-stapled ribosomes associate with endogenous ribosomal subunits to substantially different extents (Fig. 2c–e and Extended Data Fig. 3). Previously described O-ribosomes with linked subunits have cross-assembly coefficients close to one, demonstrating that they interact extensively and stably with endogenous ribosome subunits *in trans*. By contrast, O-d2d8 has substantially reduced cross-assembly coefficients (Fig. 2d, e and Extended Data Fig. 3).

Next we compared the relative activity of different O-stapled ribosomes, resulting solely from linked subunits acting in *cis*- or

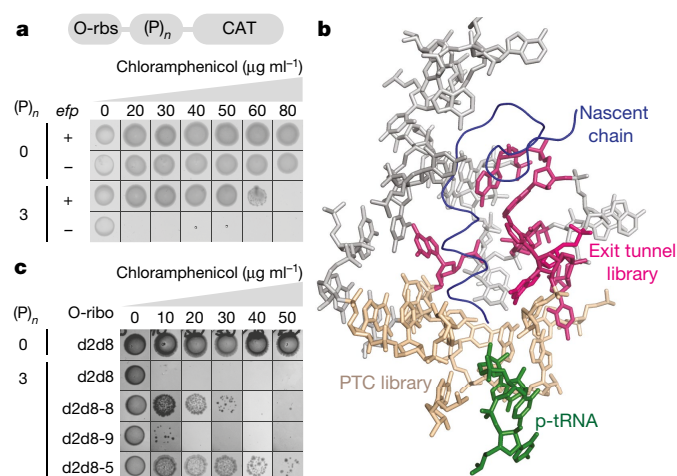


Fig. 4 | Discovering O-d2d8 variants with the intrinsic ability to translate polypyrroline sequences. **a**, Top, the O-rbs-(P)_n-CAT gene, in which an orthogonal ribosome-binding site (O-rbs) directs the translation of proline (P) codons ($n = 0$ or 3) followed by the chloramphenicol acetyltransferase (CAT) gene. When the number of proline codons is 0 , the O-d2d8 ribosome can translate the CAT gene even in the absence of *efp* (encoding EF-P), and thus confers resistance to chloramphenicol. In the absence of *efp*, the ribosome cannot translate through the polypyrroline ($n = 3$) sequence and thus cannot translate the CAT gene; hence the strain is not resistant to chloramphenicol and does not grow. The experiment was performed twice with similar results. **b**, The region of the ribosome that we targeted for mutation (PDB accession code 5NWY) in order to facilitate translation of polypyrroline sequences. Beige, nucleotides (C2063, G2447, A2450, A2451, C2452, U2506, G2583, U2584, U2585 and A2602) of the 23S rRNA in the peptidyltransferase centre (PTC) that were randomized. Pink, nucleotides (2058A, 2059A, 2061G, 2062A, 2501C, 2503G and 2505G) in the exit tunnel that were randomized. **c**, Chloramphenicol resistance provided by evolved O-d2d8 variants in the absence of EF-P. The experiment was performed twice with similar results. Note that cells grow more slowly in strains lacking EF-P.

trans-linked complexes (Fig. 1b), in the presence of endogenous subunits. We achieved this by first developing an *in vitro* orthogonal translation system (Extended Data Fig. 4 and Supplementary Data 3) in which we could selectively inhibit translational contributions from wild-type subunits by adding the antibiotics spectinomycin and erythromycin; we used stapled subunits that contain mutations conferring resistance to these antibiotics^{16,17} (Fig. 2f, Extended Data Fig. 5 and Supplementary Data 4). Several O-stapled ribosomes that translate an O-mRNA in the presence of functional endogenous subunits could not support robust translation when the endogenous subunits were inactivated (Fig. 2g, Extended Data Fig. 6); these ribosomes presumably mediate translation via *trans* interactions with endogenous subunits. By contrast, other O-stapled ribosomes—notably O-d2d8—exhibit substantial activity when the translational capacity of endogenous subunits is inhibited (Fig. 2g), and can mediate robust translation using only O-stapled ribosome subunits.

Taken together, our data suggest that the highly active translation of O-mRNAs by O-d2d8 may be mediated primarily by the association of subunits within the O-stapled ribosome, and indicate that the connection between the J5–J5a RNA hinge and the two subunits in the d2d8 linker has an optimum geometry for specifying intramolecular subunit association.

We used a variant of replicon excision enhanced recombination (REXER¹⁸; Fig. 3a) to create bacterial strains that use genomically encoded stapled ribosomes with a wild-type ASD as the sole ribosome in the cell¹⁹ (Extended Data Fig. 7). Sucrose gradient analyses of the resulting strains suggest that, in some strains, *trans*-linked stapled ribosome complexes dominate and probably support growth (Fig. 3b and Extended Data Fig. 7e). The strain containing the d2d8 stapled ribosome ($\Delta 6$ d2d8), however, forms minimal *trans*-linked complexes

(Fig. 3b), and the growth rate of the strain is 30–40% that of wild-type controls (Extended Data Fig. 7d).

We developed an automated parallel evolution method (Fig. 3c) and used it to select faster-growing strains derived from the $\Delta 6$ d2d8 strain; these strains facilitated isolation of d2d8 ribosomes for structural studies. Following strain evolution, we isolated individual clones from the four fastest-growing cultures to analyse in detail. The specific growth rates of the evolved strains were approximately 81% of those of controls with wild-type ribosomes (Fig. 3d). Whole-genome sequencing of independently evolved strains provided further insight into the mutations associated with their improvement (Supplementary Data 5–11). We note that our approach provides a rapid and reproducible route to cellular evolution (with the entire process taking just two to three weeks).

To reveal how the staple connects the ribosome subunits of d2d8 to mediate their selective association, we prepared d2d8 70S ribosomes from $\Delta 6$ d2d8-E10 *E. coli* (Extended Data Fig. 8a) and visualized the stapled ribosome structure by single-particle cryo-EM, at an overall resolution of 3.0 Å (Fig. 3e, f, Extended Data Fig. 8b, c and Extended Data Table 1). The structure clearly reveals the two subunits of the ribosome in a canonical conformation, with protein and RNA components in the same positions as in previous structures of native ribosomes²⁰. Notably, the two subunits are linked by continuous electron density between helix 44 in the small subunit and Helix 101 in the large subunit (Fig. 3e–h). A previously solved structure of the J5–J5a RNA hinge¹¹ was docked easily into this density (Fig. 3g, h); this is consistent with the hinge adopting a broadly similar structure, when stapling ribosome subunits, as in its native context. The structure reveals how the rational union of RNA modules and combinatorial optimization can be used to engineer a megadalton-sized molecular machine with new properties.

To explicitly demonstrate the advantages of controlling subunit association in the O-stapled ribosome, we evolved the large subunit of O-d2d8 to access a function that has not been accessed in the native ribosome—the intrinsic ability to polymerize a polypyrroline sequence. The natural ribosome, which is intrinsically able to translate the other 19 amino acids, commonly fails on polypyrroline (poly(P)) sequences.

The challenge of translating polypyrroline sequences arises from poor accommodation of the prolyl-tRNA, the retarded rate of peptide-bond formation at proline residues, and clashes between the nascent polypyrroline chain and the exit tunnel (notably residue G2061)^{21–23}. Nature addresses this challenge by augmenting translation with elongation factor P (EF-P)^{24,25}, which binds to the E-site of ribosomes and stabilizes and positions the peptidyl-tRNA to favour both the formation of peptide bonds between proline residues and the elongation of polypyrroline sequences. We investigated whether we could evolve O-stapled ribosomes that are intrinsically able to translate proline-rich sequences in the absence of EF-P.

We found that O-d2d8, like the natural ribosome, struggles to translate proline-rich sequences in the absence of EF-P (Fig. 4a and Extended Data Fig. 9). We then created a library that randomized ten nucleotides in the peptidyltransferase centre (PTC) of O-d2d8 (Fig. 4b), and selected those library members that can read through proline-coding stretches between an orthogonal ribosome-binding site and a chloramphenicol acetyltransferase gene (O-(P)₂-CAT and O-(P)₃-CAT) in the absence of EF-P (Fig. 4b). We identified four O-d2d8 variants (Extended Data Fig. 9c and Supplementary Data 12), but none of these constituted a general solution to translating proline-rich sequences. We next created a second library that randomized seven exit-tunnel nucleotides (Fig. 4b), using the four previously selected PTC variants as templates. We identified ten O-d2d8 variants following this selection, including O-d2d8 (5) (Supplementary Data 12). Notably, O-d2d8 (5) confers the ability to translate proline-rich sequences at a level approaching that facilitated by EF-P (Fig. 4c), enhances the translation of polypyrroline sequences of varying lengths in several contexts, and can act synergistically with EF-P (Extended Data Fig. 9d–g). O-d2d8 (5) produces (P)₇-tagged green fluorescent protein (GFP) with

the expected total mass, confirming that it translates through polyproline sequences with good fidelity (Extended Data Fig. 9h). O-d2d8 (5) contains 15 nucleotide mutations (Supplementary Data 12); the G2061U mutation and other purine-to-pyrimidine mutations may relieve the steric clash between the exit tunnel and proline residues in the growing nascent chain²¹. Our results provide a first example of accessing new, previously inaccessible function in the large subunit of an orthogonal ribosome.

We have demonstrated how controlling the association of ribosome subunits and directing both subunits to an orthogonal message enables the evolution of new large-subunit function that has not been accessed in natural ribosomes. Our work provides a foundation for further evolution of the rRNA in the entire O-stapled ribosome; it may facilitate the evolution of O-stapled ribosomes for non-natural bond-forming reactions^{26–29}, and the selective recruitment of tRNAs and other translation factors to write orthogonal genetic codes^{2,3,30}. Thus we anticipate that our approach will facilitate the encoded cellular synthesis of biopolymers with non-natural backbone chemistries⁷.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0773-z>.

Received: 27 June 2018; Accepted: 15 October 2018;

Published online 5 December 2018.

- Rackham, O. & Chin, J. W. A network of orthogonal ribosome-mRNA pairs. *Nat. Chem. Biol.* **1**, 159–166 (2005).
- Wang, K., Neumann, H., Peak-Chew, S. Y. & Chin, J. W. Evolved orthogonal ribosomes enhance the efficiency of synthetic genetic code expansion. *Nat. Biotechnol.* **25**, 770–777 (2007).
- Neumann, H., Wang, K., Davis, L., Garcia-Alai, M. & Chin, J. W. Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. *Nature* **464**, 441–444 (2010).
- Wang, K. et al. Optimized orthogonal translation of unnatural amino acids enables spontaneous protein double-labelling and FRET. *Nat. Chem.* **6**, 393–403 (2014).
- Fried, S. D., Schmied, W. H., Uttamapinant, C. & Chin, J. W. Ribosome subunit stapling for orthogonal translation in *E. coli*. *Angew. Chem. Int. Edn* **54**, 12791–12794 (2015).
- Orelle, C. et al. Protein synthesis by ribosomes with tethered subunits. *Nature* **524**, 119–124 (2015).
- Chin, J. W. Expanding and reprogramming the genetic code. *Nature* **550**, 53–60 (2017).
- Voorhees, R. M. & Ramakrishnan, V. Structural basis of the translational elongation cycle. *Annu. Rev. Biochem.* **82**, 203–236 (2013).
- Triman, K. L., Peister, A. & Goel, R. A. Expanded versions of the 16S and 23S ribosomal RNA mutation databases (16SMDbexp and 23SMDbexp). *Nucleic Acids Res.* **26**, 280–284 (1998).
- Kitahara, K. & Suzuki, T. The ordered transcription of RNA domains is not essential for ribosome biogenesis in *Escherichia coli*. *Mol. Cell* **34**, 760–766 (2009).
- Szewczak, A. A. & Cech, T. R. An RNA internal loop acts as a hinge to facilitate ribozyme folding and catalysis. *RNA* **3**, 838–849 (1997).
- Nakatogawa, H. & Ito, K. The ribosomal exit tunnel functions as a discriminating gate. *Cell* **108**, 629–636 (2002).
- Vázquez-Laslop, N., Ramu, H., Klepacki, D., Kannan, K. & Mankin, A. S. The key function of a conserved and modified rRNA residue in the ribosomal response to the nascent peptide. *EMBO J.* **29**, 3108–3117 (2010).
- Barrett, O. P. & Chin, J. W. Evolved orthogonal ribosome purification for in vitro characterization. *Nucleic Acids Res.* **38**, 2682–2691 (2010).
- Youngman, E. M. & Green, R. Affinity purification of *in vivo*-assembled ribosomes for in vitro biochemical analysis. *Methods* **36**, 305–312 (2005).
- Vester, B. & Douthwaite, S. Macrolide resistance conferred by base substitutions in 23S rRNA. *Antimicrob. Agents Chemother.* **45**, 1–12 (2001).
- Sigmund, C. D., Ettayebi, M. & Morgan, E. A. Antibiotic resistance mutations in 16S and 23S ribosomal RNA genes of *Escherichia coli*. *Nucleic Acids Res.* **12**, 4653–4664 (1984).
- Wang, K. et al. Defining synonymous codon compression schemes by genome recoding. *Nature* **539**, 59–64 (2016).
- Quan, S., Skovgaard, O., McLaughlin, R. E., Buurman, E. T. & Squires, C. L. Markerless *Escherichia coli* *rrn* deletion strains for genetic determination of ribosomal binding sites. *G3 (Bethesda)* **5**, 2555–2557 (2015).
- James, N. R., Brown, A., Gordiyenko, Y. & Ramakrishnan, V. Translational termination without a stop codon. *Science* **354**, 1437–1440 (2016).
- Huter, P. et al. Structural basis for polyproline-mediated ribosome stalling and rescue by the translation elongation factor EF-P. *Mol. Cell* **68**, 515–527 (2017).
- Doerfel, L. K. et al. Entropic contribution of elongation factor P to proline positioning at the catalytic center of the ribosome. *J. Am. Chem. Soc.* **137**, 12997–13006 (2015).
- Pavlov, M. Y. et al. Slow peptide bond formation by proline and other N-alkylamino acids in translation. *Proc. Natl Acad. Sci. USA* **106**, 50–54 (2009).
- Doerfel, L. K. et al. EF-P is essential for rapid synthesis of proteins containing consecutive proline residues. *Science* **339**, 85–88 (2013).
- Ude, S. et al. Translation elongation factor EF-P alleviates ribosome stalling at polyproline stretches. *Science* **339**, 82–85 (2013).
- Maini, R. et al. Protein synthesis with ribosomes selected for the incorporation of β -amino acids. *Biochemistry* **54**, 3694–3706 (2015).
- Melo Czekster, C., Robertson, W. E., Walker, A. S., Söll, D. & Schepartz, A. In vivo biosynthesis of a β -amino acid-containing protein. *J. Am. Chem. Soc.* **138**, 5194–5197 (2016).
- Maini, R. et al. Ribosome-mediated incorporation of dipeptides and dipeptide analogues into proteins in vitro. *J. Am. Chem. Soc.* **137**, 11206–11209 (2015).
- Dedkova, L. M., Fahmi, N. E., Golovine, S. Y. & Hecht, S. M. Construction of modified ribosomes for incorporation of D-amino acids into proteins. *Biochemistry* **45**, 15541–15551 (2006).
- Terasaka, N., Hayashi, G., Katoh, T. & Suga, H. An orthogonal ribosome-tRNA pair via engineering of the peptidyl transferase center. *Nat. Chem. Biol.* **10**, 555–557 (2014).

Acknowledgements This work was supported by the UK Medical Research Council (MRC; grants MC_U105181009 and MC_UP_A024_1008), the Biotechnology and Biological Sciences Research Council (BBSRC; grant BB/M000842/1, for automation) and the European Research Council (ERC) Advanced Grant (grant SGCR), all to J.W.C. S.D.F. was supported by a fellowship from Kings College, Cambridge. C.D.R. was supported by a Gates Cambridge Scholarship, and supported in the laboratory of V. Ramakrishnan by the MRC (grant MC_U105184332), the Wellcome Trust (grant WT096570), the Louis-Jeantet Foundation and the Agouron Institute.

Author contributions W.H.S. developed riboREXER and automated parallel evolution, and analysed the resulting data. Z.T. prepared samples for electron microscopy, developed the orthogonal in vitro translation systems and analysed the resulting data. Z.T. also generated the O-stapled ribosome library, with the assistance of W.H.S., and performed and analysed the polyproline translation experiments. C.U. developed the MS2 pulldown experiments, performed the pulldowns, and analysed the data using samples prepared with the assistance of W.H.S. C.D.R. performed cryo-EM and data analysis. S.D.F. cloned O-stapled ribosomes and performed some initial analysis. W.H.S. characterized the activities of O-stapled ribosomes, with assistance from C.U. and Z.T. W.H.S., Z.T., C.U. and J.W.C. wrote the paper, with input from all authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0773-z>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0773-z>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to J.W.C.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Generation of the linker-length library for the O-stapled ribosome. We altered the linker length in the O-stapled rDNA in a modular fashion with a modified GoldenGate cloning procedure³¹, using the primer pairs long-XX-f and long-XX-r and short-XX-f and short-XX-r, (for all oligonucleotide sequences, see Supplementary Data 1) with pRSF-O-ribo(h44H101) as a template. h44 is helix 44; H101 is Helix 101; O-ribo(h44H101) is also called O-d0d0, for consistency with the naming scheme for describing linker-length variants. All of the resulting long fragments and short fragments were uniquely combined. The resulting O-stapled rDNA plasmids were sequence verified.

In vivo activity assay for O-stapled ribosomes. *Escherichia coli* GeneHogs cells containing a plasmid carrying the *Methanosarcina mazei* pyrrolysyl-tRNA synthetase (PylS)/tRNA_{CUA} pair (pKW1³²), a reporter plasmid carrying a GFP_{150TAG} gene preceded by an orthogonal ribosome-binding site¹, and a plasmid containing a linker-length variant of the O-stapled ribosome were assayed following induction of expression with 1 mM isopropyl-β-D-1-thiogalactopyranoside (IPTG). The OD₆₀₀ and fluorescence at 485/520 nm were measured on a SpectraMax i3 plate reader (Molecular Devices). The fluorescence was normalized by OD₆₀₀ and baseline autofluorescence was deducted.

Generation of MS2-tagged O-ribosome variants. The MS2 stem loop, flanked on either end by three uridine residues and two *SpeI* sites, was inserted in place of nucleotides 83–86 at the tip of helix 6 of the 16S rDNA portion of different pRSF-O-ribo(h44H101) variants. The insertion was carried out through a one-step, site-directed mutagenesis protocol³³, using Forward-ms2 and Rev-ms2 primers.

Cloning, expression and purification of GST-tdMS2CP. The sequence encoding the bacteriophage MS2 tandem dimer (tdMS2CP) was custom-synthesized (by Genscript) to contain the V29I mutation³⁴ in one monomer, the V75E/A81G double mutation³⁵ in the other, and a GSSGGSSG linker between the two monomers. It was then subcloned into plasmid pGEX-4T1, expressed and purified on 4b glutathione-sepharose resin (GE Healthcare) that had been washed with phosphate-buffered saline (PBS). The eluent was dialysed three times in 4 l PBS supplemented with 20% v/v glycerol, aliquoted and stored at –80 °C.

Expressing MS2-tagged O-stapled ribosomes for affinity purification. DH10B cells containing the p15A-O-cat plasmid as well as a pRSF plasmid encoding a member of the MS2-tagged O-stapled-ribosome linker-length library were grown to an OD₆₀₀ 0.05, and then induced with 1 mM IPTG (3 h, 37 °C). The culture was then immediately cooled in an ice bath for 30 min to halt translation initiation and to allow ribosomes on mRNAs to finish translation. Cells from each culture were pelleted (50 ml, 4,000 r.p.m., 10 min, 4 °C), resuspended in 1 ml ribosome lysis buffer (10 mM Tris pH 8, 10 mM MgCl₂, 1 mg ml^{–1} lysozyme) and lysed for three freeze–thaw cycles; 30 μl 10% sodium deoxycholate was then added to complete lysis. The lysate was clarified by centrifugation (13,000 r.p.m., 20 min, 4 °C) and the supernatant was used for immediate analysis or was flash-frozen and stored at –80 °C.

Sucrose gradient and affinity purification of O-stapled ribosomes. Lysates containing MS2-tagged O-stapled ribosomes were first purified using sucrose gradient fractionation: 10–40% linear sucrose gradients (20 mM Tris pH 7.5, 10 mM MgCl₂, 100 mM NH₄Cl, 6 mM β-mercaptoethanol (BME), 100 μM benzamide, 100 μM phenylmethylsulfonyl fluoride (PMSF)) were generated using a BioComp Gradient Station (BioComp Instruments). Ten A₂₆₀ units of the lysate were layered directly onto each sucrose gradient (14 ml) and spun in an ultracentrifuge using a SW 40 Ti rotor at 37,000 r.p.m. for 3 h at 4 °C. Gradients were fractionated using the BioComp Gradient Station with a flow rate of 0.3 mm s^{–1}. Fractions were monitored via the absorption profile at 254 nm, and manually collected over the collection window.

To prevent nonspecific sticking, the GST-tdMS2CP-bound resins (roughly 0.5 ml solid beads) were incubated with 1 mg sheared salmon sperm DNA for 30 min at 4 °C. The pooled ribosome fractions as described above (from ten A₂₆₀ units of the lysate) were then added to the GST-tdMS2CP-bound resins (roughly 0.5 ml solid beads) and incubated with gentle mixing for 1 h at 4 °C. The resins were washed extensively, twice with 10 ml ribosome-binding buffer at 4 °C, and twice more with 10 ml ribosome-binding buffer with 0.05% Tween-20; the wash steps were separated by gentle mixing at 4 °C for at least 20 min. To harvest RNA species bound to resins, we treated the resulting resin pellet with QIAzol (500 μl) for 5 min at room temperature. Chloroform (100 μl) was then added, and the mixture vortexed and centrifuged (12,000g, 5 min, 4 °C). The upper aqueous phase was collected and mixed in a 1:1 v/v ratio with isopropanol. After incubation at –20 °C for 30 min, the RNA pellet was compacted by centrifugation (15,000g, 15 min, 4 °C), washed once with 70% ethanol, and air-dried, and the ratio of O-stapled ribosome to endogenous subunits was determined. Control experiments from cells without the MS2-tagged ribosome were used to remove contributions from nonspecific capture. We defined the 30S and 50S cross-assembly coefficients as the ratio of 30S or 50S to O-stapled ribosome. For a small number of O-stapled

ribosomes, MS2 purifications were less efficient and we were not able to determine cross-assembly coefficients.

Constructs and O-ribosomes for orthogonal translation in vitro. The construct pCF-sfGFP contains a T7 promoter, a wild-type ribosome-binding site (rbs) and an sfGFP gene. pCF-O-sfGFP has an orthogonal ribosome-binding site replacing the wild-type site, as well as an AT-rich sequence separating the T7 promoter and the O-rbs, introduced with primers z71 and z72. We introduced antibiotic-resistance mutations to the O-stapled ribosome subunits by standard polymerase chain reaction (PCR)-based methods.

Preparation of cell extracts containing orthogonal ribosomes. *E. coli* BL21 cells were transformed with pRSF-O-ribo(h44H101) variants or pRSF-O-Ribo¹. Overnight cultures were used to inoculate 2 × TYPG medium supplemented with 10 mM MgSO₄, 10 mM MgCl₂, 25 μg ml^{–1} kanamycin and 0.2 mM IPTG to a starting OD₆₀₀ of 0.05, and grown until the OD₆₀₀ reached 2.5. Cells were chilled rapidly on an ice-water bath, pelleted at a relative centrifugal force (RCF) of 5,000 at 4 °C for 10 min, and washed three times with ice-cold buffer A (10 mM Tris-acetate pH 8.2, 14 mM Mg(OAc)₂, 60 mM KOAc and 2 mM dithiothreitol (DTT)). The washed cell pellets were weighed, flash-frozen in liquid nitrogen, and stored at –80 °C. Just before use, cells were rapidly thawed and suspended in 1 ml of buffer A per gram of wet cell mass. Cells were either grown in 1 l volume in baffled 2.5 l conical flasks with agitation at 150 r.p.m. (for reference lysate samples), or in 50 ml volume in 250 ml Erlenmeyer flasks with agitation at 270 r.p.m. (for activity screening of O-stapled-ribosome variants). Cells grown on a litre scale were lysed with EmulsiFlex-C3 homogenizer (Avestin, Ottawa, Canada) and prepared as reported³⁶. Cells grown on a 50-ml scale were lysed via sonication using a VCX750 Instrument (Sonics&Materials, USA) equipped with a 2-mm-diameter 8-tip probe (QSonica, USA) as described³⁶, with the following sonication parameters: frequency 20 kHz, 20% of amplitude, 5 s/10 s sonication/rest regime, and energy input 400 J. Lysates were clarified via centrifugation at 21,000 RCF at 4 °C for 20 min. The top 200 μl of the lysate were taken and flash-frozen in liquid nitrogen, then stored at –80 °C. The total amount of protein in cell extracts, as quantified by the Bradford detergent-free protein assay, was consistently 8 ± 2.8 mg ml^{–1}.

Translation by O-ribosomes after native-ribosome inhibition. In vitro translation reactions were carried out in a 384-well plate with reaction volumes of 12.5 μl. The reaction mixture for protein synthesis was prepared as described³⁷. Reactions were carried out at 30 °C for at least 4 h. Synthesis of superfolder GFP (sfGFP) was monitored on an Infinite 200 Pro plate reader (Tecan AG, Switzerland).

We carried out three types of in vitro translation reaction for each O-stapled ribosome variant. First, to control for the quality of the lysate that contains a mixture of endogenous and O-stapled ribosomes, we added the reporter plasmid pCF-sfGFP (which includes the wild-type ribosome-binding site) to the reaction and measured GFP production from the orthogonal ribosome-binding site. Finally, we added pCF-O-sfGFP, 10 μM spectinomycin and 50 μM erythromycin and measured GFP production from the orthogonal ribosome-binding site when native subunits were inhibited. The GFP signal obtained from the reaction with pCF-sfGFP was used to normalize the GFP signal measured from pCF-O-sfGFP. The background signal that originates from the activity of endogenous ribosomes on the orthogonal message was measured in lysates containing only endogenous ribosomes and pCF-O-sfGFP, and was subtracted from the normalized GFP signals measured when an O-stapled ribosome was present in the lysate. The activity of an O-stapled ribosome in the presence of 10 μM spectinomycin and 50 μM erythromycin divided by its activity in the absence of antibiotics provides a ratio that allows us to compare O-ribosomes of different total activity. For graphing, the highest such ratio was set to 1 and relative values between 0 and 1 are reported for other O-stapled ribosomes. We cannot determine the absolute fractional activity of translation resulting from only O-ribosome subunits because we cannot determine the extent to which erythromycin inhibits the stapled large subunit. Previously described ribosomes with linked subunits had low in vitro translation activities that precluded their characterization in this system.

Genomic replacement of an rRNA operon by ribo-REXER. Ribo-REXER is a modification of REXER¹⁸. A landing site—containing the chloramphenicol acetyltransferase (CAT) gene as a positive and the *SacB* gene as a negative selection marker—was introduced at the 5' flank of the *rrnE* rRNA operon in the genome of SQ110 *E. coli*¹⁹ by Red/ET recombination (Gene Bridges). Correct integration was validated by phenotyping, colony PCR using primer pair WS441f/r, and sequencing, and resulted in plasmid SQ110^{CAT-SacB}.

Candidate CRISPR target sequences 5' and 3' of the *rrnE* region were identified using ChopChop^{38,39} and matched to sequences on the incoming donor plasmid, so that the linearized product would retain homology arms of 50 base pairs on both the 5' and the 3' flanks with respect to the targeted genomic region. The plasmid containing the four spacers for REXER 4 was cloned by PCR mutagenesis of the spacer plasmid¹⁸, and subsequently shortened to contain only the two-spacer cassette necessary for REXER 2.

The donor plasmid containing flanking homology regions and protospacer adjacent motif (PAM) sequences for integration at the *rrnE* locus was constructed by In-Fusion cloning (Takara). The sequences chosen for targeting 5' of the incoming rRNA operon and 3' of the pH cassette (which comprises the *E. coli* PheS-derived negative selection marker PheS* (T251A A294G)⁴⁰ paired with an *hph* gene, enabling resistance to hygromycin B) were CATTAATTGCGTTGCGCACGGGG and AGGCAAGACCGAGCGCCATTGG. Donor plasmids containing stapled O-rDNA were created from the parent donor plasmid.

REXER was performed essentially as described¹⁸ using the spacer cassette:

taataactaaaatgtgataatactcttaataatgcagtaacaggggctttcaagactgaagtctagctga gacaatagctgctgattacgaaatgttttagacaaaatagctacgaggttttagagctatgctgttgatgtcccaaacCattaatgctgtgcgcacgggttttagagctatgctgtttgaatgtcccaaacAGGCAAGACCGAG CGCCATTGtttagagctatgctgtttgaatgtcccaaacctcagcacactgagactgttgagttgaattcg gtcagtcgcg.

We identified rDNA replacements by colony PCR using primers WS467f/341r and WS218f/467r, and total RNA extraction. Loss of the incoming plasmid was confirmed by extracting plasmid DNA (Plasmid Miniprep Kit, Qiagen) and testing (by PCR amplification of the *SpcR* locus) for the presence of the tRNA helper plasmid pTrNA67, present in the $\Delta 6$ and $\Delta 7$ strains, and for loss of the donor-plasmid backbone, and also by transforming TOP10 *E. coli* cells with the extracted plasmid mixture and testing for the formation of colonies on LB agar plates containing 75 $\mu\text{g ml}^{-1}$ spectinomycin, but not 150 $\mu\text{g ml}^{-1}$ hygromycin B. Growth rates were determined by growing at minimum eight colonies from a validated strain in $2 \times \text{TY}$ growth medium containing 75 $\mu\text{g ml}^{-1}$ spectinomycin, 150 $\mu\text{g ml}^{-1}$ hygromycin B and 1% glucose.

Automated parallel evolution. The $\Delta 6$ d2d8 strain was transformed with the mutagenesis plasmid MP6 (ref. ⁴¹) and grown on LB agar containing 2% glucose, 150 $\mu\text{g ml}^{-1}$ hygromycin B, 75 $\mu\text{g ml}^{-1}$ spectinomycin and 50 $\mu\text{g ml}^{-1}$ chloramphenicol. We picked 92 individual colonies into wells containing liquid media. Random mutagenesis was induced by addition of 25 mM arabinose to each well, creating mutator cultures in each well. The mutator cultures were incubated for 48 h at 37°C and 200 r.p.m.⁴¹. Following a 1/400 dilution step into fresh growth media, the incubation was repeated. The remaining four wells of the culture plate were inoculated with colonies from $\Delta 6$ d2d8 and grown in 1% glucose.

We diluted 0.5 μl of each culture into 200 μl of liquid medium containing 1% glucose in a fresh lidded 'growth' plate, and then incubated the plates (37°C, 400 r.p.m., 2 h). We measured the OD₆₀₀ in each well at the start and then again periodically (every 2–9 h) depending on the growth rate. When the OD₆₀₀ reached a threshold of 0.5 for at least three mutator-culture wells, all cultures were diluted. If fewer wells reached the threshold, the three wells with the present highest OD₆₀₀ were identified. This process was repeated once more, and if the threshold was not reached after 20 h, the plate was incubated for a further 12 h before the dilution step was implemented independently of the measured cell density. This allowed for dynamic maintenance of a week-long evolution experiment that flexibly adjusted the selective pressure on the system on the basis of the observed bacterial growth behaviour during each step. After an estimated 110 generations, the workflow—which we implemented on a Beckmann robotics platform—was terminated and glycerol stocks were created from the final growth plate. The pools with the highest final OD₆₀₀ measurements were carried forward, and growth curves were recorded for individual clonal lines derived from each pool.

The resulting $\Delta 6$ d2d8 strains were characterized by colony PCR and total RNA extraction, and their genomes were sequenced (Miseq, Illumina). The sequencing data were aligned to a reference sequence derived from strain SQ110 (GenBank, NCBI) using a Bowtie script (<https://github.com/TiongSun/iSeq>)⁴². Every called mutation was then filtered⁴³ to identify mutations introducing an in-frame stop codon and any non-silent mutation within an open reading frame.

Stapled ribosome preparation for cryo-EM. *Escherichia coli* cells expressing only stapled ribosomes ($\Delta 6$ d2d8 E10) were grown in 6 l of $2 \times \text{TYPG}$ media supplemented with 5 mM MgCl₂ at 37°C with agitation, until the OD₆₀₀ reached 0.5–0.6. Cells were rapidly chilled in an ice-water bath, harvested by centrifuging at 5,000 RCF at 4°C for 10 min, washed twice in ice-cold TBST buffer and washed once with ice-cold buffer '100/10' (20 mM HEPES-KOH pH 7.6, 100 mM NH₄Cl, 10.5 mM Mg(OAc)₂, 0.5 mM EDTA, 2 mM DTT). The cell pellet was resuspended in ten volumes of ice-cold '100/10' buffer and lysed with EmulsiFlex-C3 homogenizer (Avestin, Ottawa, Canada) at a variable pressure of 17,000 to 20,000 p.s.i., passing twice. Cell debris was removed by two rounds of centrifugation at 30,000 RCF at 4°C for 30 min, with only the top 60–80% of the volume being passed to the next stage. The S30 supernatant was loaded on 1.25 volumes of 1.1 M sucrose in 20 mM HEPES-KOH pH 7.6, 500 mM NH₄Cl, 10.5 mM Mg(OAc)₂, 0.5 mM EDTA, 2 mM DTT and centrifuged in a Type 45 Ti rotor (Beckman-Coulter) for 18 h at 37,000 r.p.m. at 4°C. Glassy pellets were washed three times with ice-cold '100/10' buffer and dissolved in the same buffer by agitation on ice. Then, approximately 400 OU₂₆₀ (where OU is optical unit) of salt-washed ribosomes were loaded onto six SW32 tubes containing 10–40%

linear sucrose gradients in '100/10' buffer and centrifuged for 18 h at 17,000 r.p.m. and 4°C. Fractions corresponding to the 70S peak were collected on a BioComp gradient station, pelleted in a Type 70 Ti rotor (Beckman-Coulter) at 55,000 r.p.m. for 2.5 h and 4°C, and dissolved in '100/10' buffer. The sample concentration was estimated at 485 OU₂₆₀ per ml and aliquots were flash-frozen in liquid nitrogen.

Cryo-electron microscopy. Quantifoil copper R2/2 400-mesh grids were coated with a thin sheet (around 60 Å) of amorphous carbon and glow-discharged for 5 s at 5 mA. Purified *E. coli* 70S d2d8 ribosomes were diluted to 100 nM in 50 mM HEPES pH 7.4, 100 mM KOAc and 5 mM Mg(OAc)₂ and applied to grids in 3- μl aliquots. Grids were incubated for 30 s at 4°C and 100% humidity, blotted for 4.5 s, and frozen in liquid ethane using a VitroBot Mark III (FEI). Micrograph movies of d2d8 70S ribosomes were collected on a Titan Krios microscope at 300 keV with a Falcon III detector using automated data collection with EPU software (all FEI). Movies were collected at a pixel size of 1.06 Å with a dose rate of 15 e[−] Å^{−2} s^{−1} over a 1.79-s exposure, consisting of 71 total frames. Defocus values of −3.2, −2.9, −2.6, −2.3, −2.0 and −1.7 μm were used.

Image processing. All processing was done using RELION-2.1 (ref. ⁴⁴). Micrograph movie frames were aligned using Motioncorr⁴⁵ and contrast transfer functions calculated using Gctf⁴⁶. Aligned movies were removed after manual inspection if micrographs contained ice particles or if contrast transfer functions failed to calculate. Ribosome particles were picked semi-autonomously⁴⁷; incorrectly picked non-ribosomal particles were identified and discarded by reference-free two-dimensional classification. The resulting 306,214 particles were used for initial three-dimensional refinement with an *E. coli* 70S ribosome (EMD-3493) low-pass-filtered to 40 Å as a reference. Three-dimensional classification of the initial reconstruction was performed without alignments to discard poorly aligned particles. Two major classes were obtained, one of closed 70S ribosomes (94,461 particles) and another containing 'opened' 70S ribosomes (128,436 particles). The closed ribosomes were selected and refined. Focused classification with signal subtraction⁴⁸ on the RNA hinge gave the final primary class (94,371 particles). The quality of the density in the hinge region is lower than that in other parts of the structure. However, classification focused on the staple revealed a single class, demonstrating sample homogeneity. This indicated that the lower density in this region is not due to a mixture of particles; it may instead reflect flexibility in the hinge and/or local variation in the hinge conformation.

Model building. A model (PDB 5MDZ) of the *E. coli* 70S ribosome²⁰ was docked into the reconstruction in Chimera⁴⁹, and individual RNA and protein chains were rigid-body-fitted using Coot⁵⁰. Portions of helix 44 from the 16S rRNA and Helix 101 from the 23S rRNA were deleted according to the sequence of the d2d8 ribosome. The RNA hinge expected to link the two subunits was then docked into the remaining, unaccounted-for density. Real-space refinement was carried out using Phenix⁵¹ and the model was validated using MolProbity⁵². Figures were created using Pymol⁵³ or Chimera⁴⁹.

O-stapled ribosomes for polypyrrole polymerization. A PTC library with a theoretical diversity of 1.1×10^6 was constructed by randomizing nucleotides C2063, G2447, A2450, A2451, C2452, U2506, G2583, U2584, U2585 and A2602 of the 23S rRNA portion of O-d2d8 through rounds of enzymatic inverse PCR, with pRSF-O-d2d8 as a starting template and using primers WS297_Gf, WS297_Tf, WS297_Af and WS297r. The resulting four clones were used in parallel enzymatic inverse PCR reactions with WS275f and WS275r. The four 'sub-libraries' were combined to form the final library with all ten positions randomized. Transformations led to around 3×10^7 to 10×10^7 colony-forming units (CFU) per sub-library (theoretical diversity 2.6×10^5); this 100-fold oversampling leads to high confidence that the library's theoretical diversity is covered in the plasmids collected. Libraries were also verified by direct sequencing of the randomized DNA and by sequencing around 20 separate clones.

The exit-tunnel library was prepared by randomizing the 23S rRNA nucleotides 2058A, 2059A, 2061G, 2062A, 2501C, 2503G and 2505G from the four most active mutants in the PTC library, namely O-d2d8^{PTC}-15, O-d2d8^{PTC}-36, O-d2d8^{PTC}-50 and O-d2d8^{PTC}-53, by enzymatic inverse PCR. The first reactions were performed with primers Lib_1_15, Lib_1_36_50 or Lib_1_53 and a Lib1r_all. The second enzymatic inverse PCR used Lib2_r_all, as well as one of three forward primers, Lib_2_15, Lib_2_36_50 or Lib_2_53. The transformation efficiencies at both stages of library preparation were around 10^7 CFU μg^{-1} of plasmid DNA. Before transformation into the ribosome-selection strain, all four sub-libraries were mixed in equal proportions. Transformations resulted in approximately 10^7 CFU per sub-library (theoretical diversity 6.6×10^4); thus, oversampling ensures that the library's theoretical diversity is covered in the plasmids collected. Libraries were also verified by direct sequencing of the randomized DNA and by sequencing of 12 clones from each sub-library.

To delete *eff* in TOP10 cells, the promoter and *hph* gene conferring resistance to hygromycin B were amplified from the pH plasmid using primer pairs z139/z140 and z141/z142, respectively, fused in an overlapping PCR with z139 and z142 primers. Primers F1/RV1 and F2/RV2 were used to introduce homology regions

and FRT recombination sites. Column-cleaned PCR product was used for lambda red recombination⁵⁴. The correct insertion of the *hph* gene at the *efp* locus was verified in hygromycin-resistant colonies by colony PCR reactions using primers z128/z144 and z128/z129.

p15A-O-(P)_n-CAT was constructed from a plasmid derived from p15-O-CAT with the forward primer z150 and the reverse primers z145, z146, z147, z148 and z149. p15A-O-(P)₄-GFP and p15A-O-(P)₇-GFP were cloned by replacing the CAT gene in the corresponding p15A-O-(P)₄-CAT and p15A-O-(P)₇-CAT plasmids with the sfGFP gene using Gibson assembly reactions.

To select O-d2d8 ribosome variants capable of translating polyproline sequences, we transformed *E. coli* TOP10 Δ efp cells with p15A-(P)₂-O-CAT or p15A-(P)₃-O-CAT reporter plasmids. The pRSF-O-d2d8 libraries were transformed into these cells by electroporation⁵⁵. Following growth and then induction of rRNA expression (1 mM IPTG, 4 h), the equivalent of 1 ml of culture at an OD₆₀₀ of 1 was spread out on 2 × TY agar 20 cm × 20 cm plates with 12.5 μg ml⁻¹ kanamycin, 4 μg ml⁻¹ tetracycline, 1 mM IPTG and 20 μg ml⁻¹ or 50 μg ml⁻¹ chloramphenicol. The plates were incubated at 37 °C for 4 days. Following selection, pRSF-O-d2d8 mutant plasmids were isolated and retransformed into *E. coli* TOP10 Δ efp cells containing the reporter plasmid p15A-(P)₂-O-CAT or p15A-(P)₃-O-CAT and used to check the phenotype. To ensure that mutations in the 23S rRNA portion of O-d2d8 were responsible for the observed phenotype, we subcloned the selected 23S, PTC-containing fragments into a fresh pRSF-O-d2d8 backbone. All obtained recloned mutant plasmids were sequence verified. To test the activity of O-d2d8 mutants in translating O-(P)_n-GFP reporters, we transformed Δ efp *E. coli* cells containing p15A-(P)_n-sfGFP reporters with pRSF-O-d2d8 mutant plasmids. GFP fluorescence and cell density were measured hourly on a SpectraMax i3 Multi-Mode Detection Platform (Molecular Devices) in a culture induced with IPTG. The measured GFP fluorescence was normalized by cell density (OD₆₀₀).

Statistics for figures. For Fig. 2b, the number of replicates (*n*) was 6 for all data, except in the case of p3d5, where *n* = 5. Error bars represent ± s.e.m.

For Fig. 3d and Extended Data Fig. 7c, growth rates were determined from *n* = 8 independently grown bacterial colonies. Curve fitting (logistic growth) was performed using Prism 7 (Graphpad). Error bars show ± s.e.m. It is not possible to plot individual points because the means and errors were derived from the best fit to all growth curves simultaneously.

For Extended Data Fig. 2a, *n* = 6 for all MS2 experiments. The standard deviations for MS2-tagged O-stapled ribosome variants were as follows (in parentheses): d0d0 (12.4), d0d1 (23.7), d0d2 (25.5), d0d3 (21.5), d0d4 (20.9), d0d5 (34.4), d0d6 (39.1), d0d7 (35.4), d0d8 (31.0), d1d0 (15.0), d1d1 (18.6), d1d2 (23.0), d1d3 (22.5), d1d4 (1.6), d1d5 (31.7), d1d6 (28.7), d1d7 (33.0), d1d8 (25.1), d2d0 (33.4), d2d1 (36.0), d2d2 (28.2), d2d3 (29.7), d2d4 (21.7), d2d5 (27.1), d2d6 (11.2), d2d7 (35.0), d2d8 (36.7), d3d0 (25.0), d3d1 (36.9), d3d2 (30.9), d3d3 (27.4), d3d4 (19.3), d3d5 (31.6), d3d6 (22.2), d3d7 (30.3), d3d8 (45.0), d4d0 (14.7), d4d1 (17.1), d4d2 (12.9), d4d3 (2.0), d4d4 (7.6), d4d5 (24.3), d4d6 (23.4), d4d7 (29.8), d4d8 (32.9), d5d0 (8.6), d5d1 (16.5), d5d2 (24.1), d5d3 (23.9), d5d4 (20.0), d5d5 (23.6), d5d6 (29.8), d5d7 (31.5), d5d8 (30.8), p1d0 (13.8), p1d1 (27.3), p1d2 (27.3), p1d3 (21.8), p1d4 (3.9), p1d5 (40.8), p1d6 (36.4), p1d7 (36.3), p1d8 (27.3), p2d0 (7.1), p2d1 (29.4), p2d2 (33.3), p2d3 (25.7), p2d4 (22.9), p2d5 (29.2), p2d6 (30.8), p2d7 (30.0), p3d0 (2.7), p3d1 (25.2), p3d2 (22.3), p3d3 (21.9), p3d4 (32.5), p3d5 (49.1), p3d6 (43.5), p3d7 (37.3) and p3d8 (12.0).

For untagged ribosomes, *n* = 6 for all data except in the case of p3d5, where *n* = 5. The standard deviations for untagged O-stapled ribosome variants were d0d0 (7.9), d0d1 (8.7), d0d2 (4.8), d0d3 (5.9), d0d4 (3.2), d0d5 (6.2), d0d6 (13.6), d0d7 (4.9), d0d8 (24.2), d1d0 (5.4), d1d1 (4.5), d1d2 (4.4), d1d3 (18.0), d1d4 (0.7), d1d5 (6.9), d1d6 (3.8), d1d7 (4.5), d1d8 (26.5), d2d0 (12.2), d2d1 (7.0), d2d2 (0.8), d2d3 (5.0), d2d4 (8.4), d2d5 (5.9), d2d6 (5.7), d2d7 (5.0), d2d8 (6.9), d3d0 (11.1), d3d1 (6.2), d3d2 (5.5), d3d3 (5.3), d3d4 (6.0), d3d5 (7.0), d3d6 (21.3), d3d7 (25.1), d3d8 (9.2), d4d0 (6.3), d4d1 (4.9), d4d2 (4.7), d4d3 (1.2), d4d4 (2.7), d4d5 (8.5), d4d6 (22.0), d4d7 (26.9), d4d8 (26.9), d5d0 (5.1), d5d1 (7.2), d5d2 (2.0), d5d3 (2.3), d5d4 (2.4), d5d5 (13.7), d5d6 (12.0), d5d7 (6.6), d5d8 (5.2), p1d0 (5.9), p1d1 (8.3), p1d2 (4.7), p1d3 (4.3), p1d4 (0.6), p1d5 (5.9), p1d6 (14.7), p1d7 (10.1), p1d8 (29.9), p2d0 (2.9), p2d1 (7.2), p2d2 (6.7), p2d3 (7.0), p2d4 (12.6), p2d5 (17.7), p2d6 (19.3), p2d7 (30.9), p3d0 (0.9), p3d1 (5.9), p3d2 (8.8), p3d3 (11.9), p3d4 (21.4), p3d5 (22.8), p3d6 (22.9), p3d7 (44.3) and p3d8 (9.3).

For Extended Data Fig. 4b, *n* values for the in vitro measurements are the same as in Extended Data Fig. 5c (below). For the in vivo measurements *n* = 6, except in the case of p3d5, where *n* = 5. Error bars represent s.e.m.

For Extended Data Fig. 4e, there were *n* = 4 independent replicates. The standard deviations are: (0, 0) ± 0.4, (5, 0) ± 0.4, (10, 0) ± 0, (20, 0) ± 0, (50, 0) ± 0, (100, 0) ± 0, (0, 5) ± 4.6, (5, 5) ± 21.8, (10, 5) ± 13.3, (20, 5) ± 7.5, (50, 5) ± 11.6, (100, 5) ± 8.3, (0, 10) ± 4.3, (5, 10) ± 14.1, (10, 10) ± 31.2, (20, 10) ± 32.6, (50, 10) ± 26.7, (100, 10) ± 16, (0, 20) ± 3.6, (5, 20) ± 37.5, (10, 20) ± 38.3, (20, 20) ± 24.8, (50, 20) ± 37.9, (100, 20) ± 23.4, (0, 50) ± 2.7, (5, 50) ± 38.7, (10,

50) ± 55.7, (20, 50) ± 30.7, (50, 50) ± 20.6, (100, 50) ± 40.1, (0, 100) ± 5.9, (5, 100) ± 20.8, (10, 100) ± 23.8, (20, 100) ± 15.3, (50, 100) ± 27.2 and (100, 100) ± 29.8, where the first digit in the parentheses corresponds to the concentration of erythromycin and the second to the concentration of spectinomycin (both in μM).

For Extended Data Fig. 5a, *n* (in parentheses) was as follows: d0d0 (3), d0d2 (3), d0d5 (4), d0d6 (4), d0d7 (5), d0d8 (4), d1d5 (3), d1d6 (3), d1d7 (8), d1d8 (4), d2d3 (4), d2d5 (4), d2d6 (4), d2d7 (4), d2d8 (4), d3d1 (3), d3d2 (4), d3d5 (4), d3d6 (4), d3d7 (4), d3d8 (4), d4d5 (4), d4d6 (4), d4d7 (10), d4d8 (10), d5d4 (3), d5d7 (4), p1d5 (4), p1d6 (4), p1d7 (3), p1d8 (4), p2d1 (4), p2d4 (3), p2d5 (3), p2d6 (5), p2d7 (3), p3d5 (4), p3d6 (3), p3d7 (3) and p3d8 (2).

For Extended Data Fig. 5b, *c*, *n* was as follows: d0d0 (3), d0d2 (3), d0d5 (4), d0d6 (4), d0d7 (4), d0d8 (4), d1d5 (3), d1d6 (3), d1d7 (7), d1d8 (4), d2d3 (4), d2d5 (4), d2d6 (3), d2d7 (4), d2d8 (4), d3d1 (3), d3d2 (4), d3d5 (3), d3d6 (4), d3d7 (4), d3d8 (4), d4d5 (4), d4d6 (4), d4d7 (8), d4d8 (7), d5d4 (2), d5d7 (4), p1d5 (4), p1d6 (3), p1d7 (2), p1d8 (4), p2d1 (4), p2d4 (3), p2d5 (2), p2d6 (4), p2d7 (3), p3d5 (4), p3d6 (3), p3d7 (2) and p3d8 (2).

For Extended Data Fig. 5d, *n* was as follows: d0d0 (5), d0d2 (5), d0d5 (4), d0d6 (3), d0d7 (3), d0d8 (3), d1d5 (4), d1d6 (4), d1d7 (5), d1d8 (4), d2d0 (4), d2d3 (4), d2d5 (4), d2d6 (4), d2d7 (3), d2d8 (4), d3d1 (4), d3d2 (4), d3d5 (4), d3d6 (4), d3d7 (5), d3d8 (5), d4d5 (5), d4d6 (3), d4d7 (4), d4d8 (5), d5d0 (4), d5d4 (4), d5d5 (6), d5d6 (6), d5d7 (4), d5d8 (4), p1d5 (5), p1d6 (4), p1d7 (3), p1d8 (3), p2d1 (3), p2d3 (4), p2d4 (3), p2d5 (3), p2d6 (4), p2d7 (3), p2d8 (3), p3d5 (4), p3d6 (3) and p3d7 (3).

For Extended Data Fig. 5e, *f*, *n* was as follows: d0d0 (4), d0d2 (3), d0d5 (4), d0d6 (5), d0d7 (5), d0d8 (4), d1d5 (3), d1d6 (3), d1d7 (5), d1d8 (4), d2d0 (4), d2d3 (4), d2d5 (4), d2d6 (4), d2d7 (4), d2d8 (4), d3d1 (3), d3d2 (4), d3d5 (4), d3d6 (4), d3d7 (4), d3d8 (4), d4d5 (5), d4d6 (5), d4d7 (5), d4d8 (5), d5d0 (3), d5d4 (5), d5d5 (4), d5d6 (4), d5d7 (4), d5d8 (4), p1d5 (4), p1d6 (4), p1d7 (4), p1d8 (4), p2d1 (4), p2d3 (3), p2d4 (3), p2d5 (3), p2d6 (4), p2d7 (3), p2d8 (3), p3d5 (4), p3d6 (3) and p3d7 (3).

For Fig. 2g and Extended Data Fig. 6, *n* was as follows: d5d7 (4), d5d6 (3), d5d5 (4), d5d4 (3), d5d0 (3), d4d8 (5), d4d7 (6), d4d6 (8), d4d5 (8), d3d8 (4), d3d7 (4), d3d6 (4), d3d5 (4), d3d2 (4), d3d1 (3), d2d8 (4), d2d7 (4), d2d6 (3), d2d5 (4), d2d3 (4), d1d8 (7), d1d7 (6), d1d6 (3), d1d5 (3), d0d8 (4), d0d7 (5), d0d6 (3), d0d5 (4), d0d2 (3), d0d0 (3), p1d8 (6), p1d7 (4), p1d6 (4), p1d5 (3), p2d8 (3), p2d7 (3), p2d6 (5), p2d5 (3), p2d4 (3), p2d3 (3), p2d1 (4), p3d8 (2), p3d7 (3), p3d6 (3) and p3d5 (4).

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

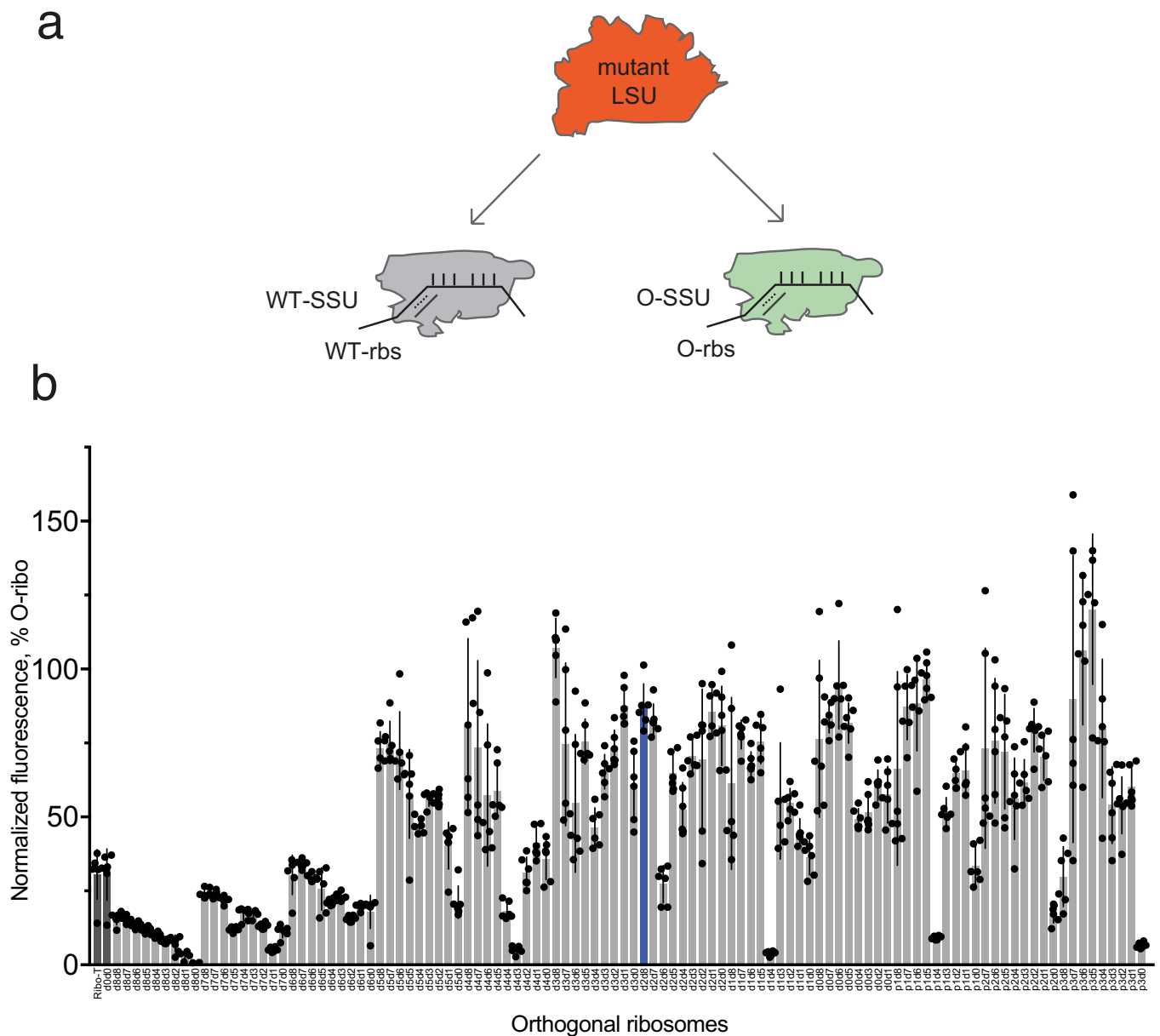
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The cryo-EM structure of d2d8 can be found under the PDB accession code 6HRM and the Electron Microscopy Data Bank accession number EMD-0261. Genome sequences for the strains created here are provided in Supplementary Tables 5–10. All other datasets generated and analysed here are available from the corresponding author upon reasonable request.

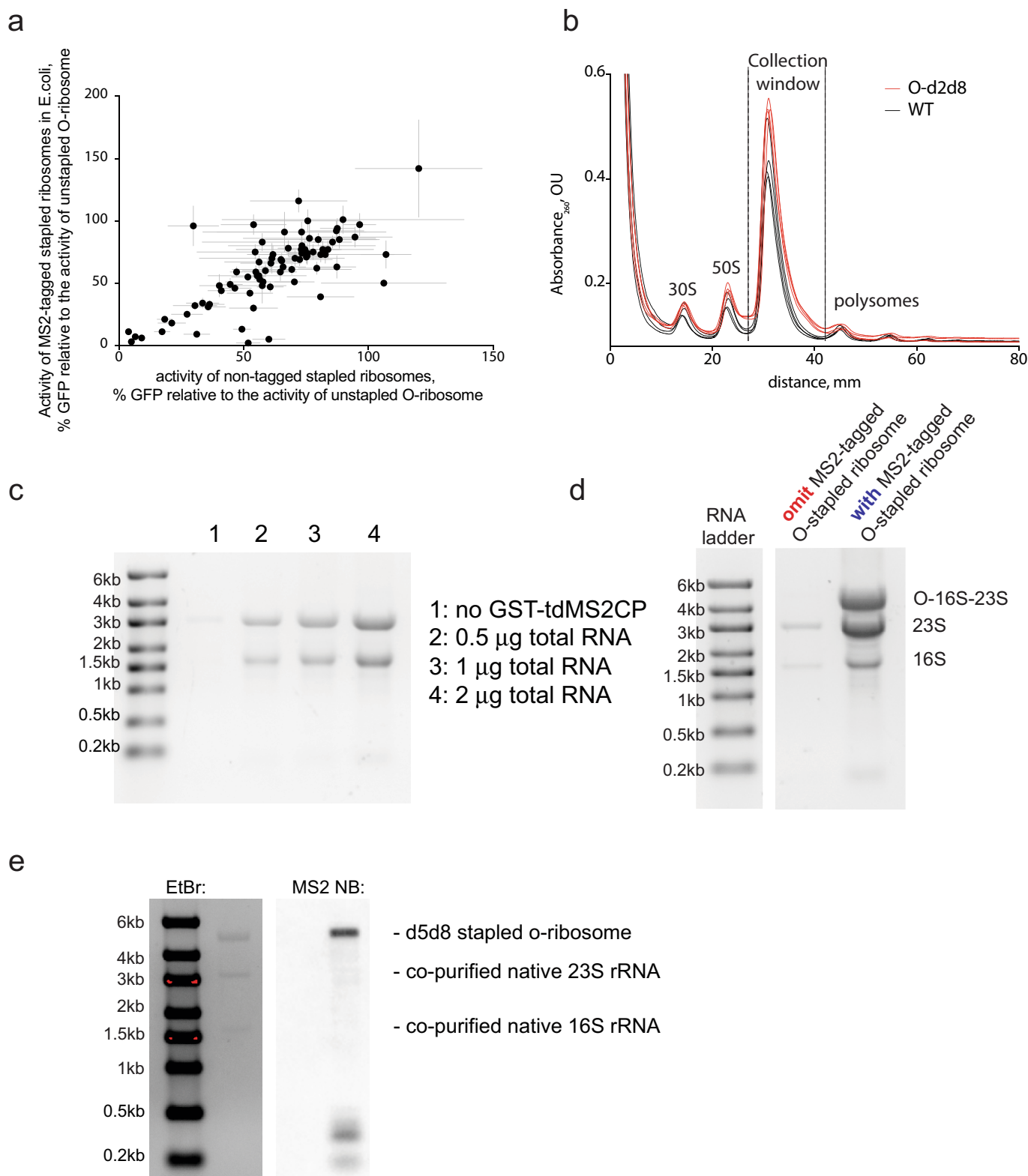
- Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS One* **3**, e3647 (2008).
- Sachdeva, A., Wang, K., Elliott, T. & Chin, J. W. Concerted, rapid, quantitative, and site-specific dual labeling of proteins. *J. Am. Chem. Soc.* **136**, 7785–7788 (2014).
- Liu, H. & Naismith, J. H. An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC Biotechnol.* **8**, 91 (2008).
- Peabody, D. S. & Ely, K. R. Control of translational repression by protein-protein interactions. *Nucleic Acids Res.* **20**, 1649–1655 (1992).
- LeCuyer, K. A., Behlen, L. S. & Uhlenbeck, O. C. Mutants of the bacteriophage MS2 coat protein that alter its cooperative binding to RNA. *Biochemistry* **34**, 10600–10606 (1995).
- Kwon, Y. C. & Jewett, M. C. High-throughput preparation methods of crude extract for robust cell-free protein synthesis. *Sci. Rep.* **5**, 8663 (2015).
- Yang, W. C., Patel, K. G., Wong, H. E. & Swartz, J. R. Simplifying and streamlining *Escherichia coli*-based cell-free protein synthesis. *Biotechnol. Prog.* **28**, 413–420 (2012).
- Labun, K., Montague, T. G., Gagnon, J. A., Thyme, S. B. & Valen, E. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res.* **44** (W1), W272–W276 (2016).
- Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M. & Valen, E. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Res.* **42**, W401–W407 (2014).
- Miyazaki, K. Molecular engineering of a PheS counterselection marker for improved operating efficiency in *Escherichia coli*. *Biotechniques* **58**, 86–88 (2015).
- Badran, A. H. & Liu, D. R. Development of potent in vivo mutagenesis plasmids with broad mutational spectra. *Nat. Commun.* **6**, 8425 (2015).

42. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
43. Cock, P. J. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
44. Fernandez-Leiro, R. & Scheres, S. H. W. A pipeline approach to single-particle processing in RELION. *Acta Crystallogr. D Struct. Biol.* **73**, 496–502 (2017).
45. Li, X. et al. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods* **10**, 584–590 (2013).
46. Zhang, K. Gctf: real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
47. Scheres, S. H. Semi-automated selection of cryo-EM particles in RELION-1.3. *J. Struct. Biol.* **189**, 114–122 (2015).
48. Bai, X. C., Rajendra, E., Yang, G., Shi, Y. & Scheres, S. H. Sampling the conformational space of the catalytic subunit of human γ -secretase. *eLife* **4**, e11182 (2015).
49. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
50. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
51. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
52. Chen, V. B., et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010).
53. The PyMOL Molecular Graphics System v.8 (Schrödinger, 2015).
54. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. USA* **97**, 6640–6645 (2000).
55. Warren, D. J. Preparation of highly efficient electrocompetent *Escherichia coli* using glycerol/mannitol density step centrifugation. *Anal. Biochem.* **413**, 206–207 (2011).



Extended Data Fig. 1 | Partitioning of free large ribosomal subunits between wild-type and orthogonal small subunits, and in vivo activity of O-stapled ribosomes. a, Gain-of-function mutations in free large subunits may confer gain-of-function phenotypes through statistical partitioning of free large subunits between wild-type (WT) and orthogonal small subunits. A mutant large subunit (LSU) can partition between WT and orthogonal (green) small subunits (SSUs) in cells that contain both WT large subunits (not shown) and mutant large subunits. Hence the mutant phenotype will be observed in the translation of both WT and orthogonal messages (see Extended Data Fig. 4d for an example).

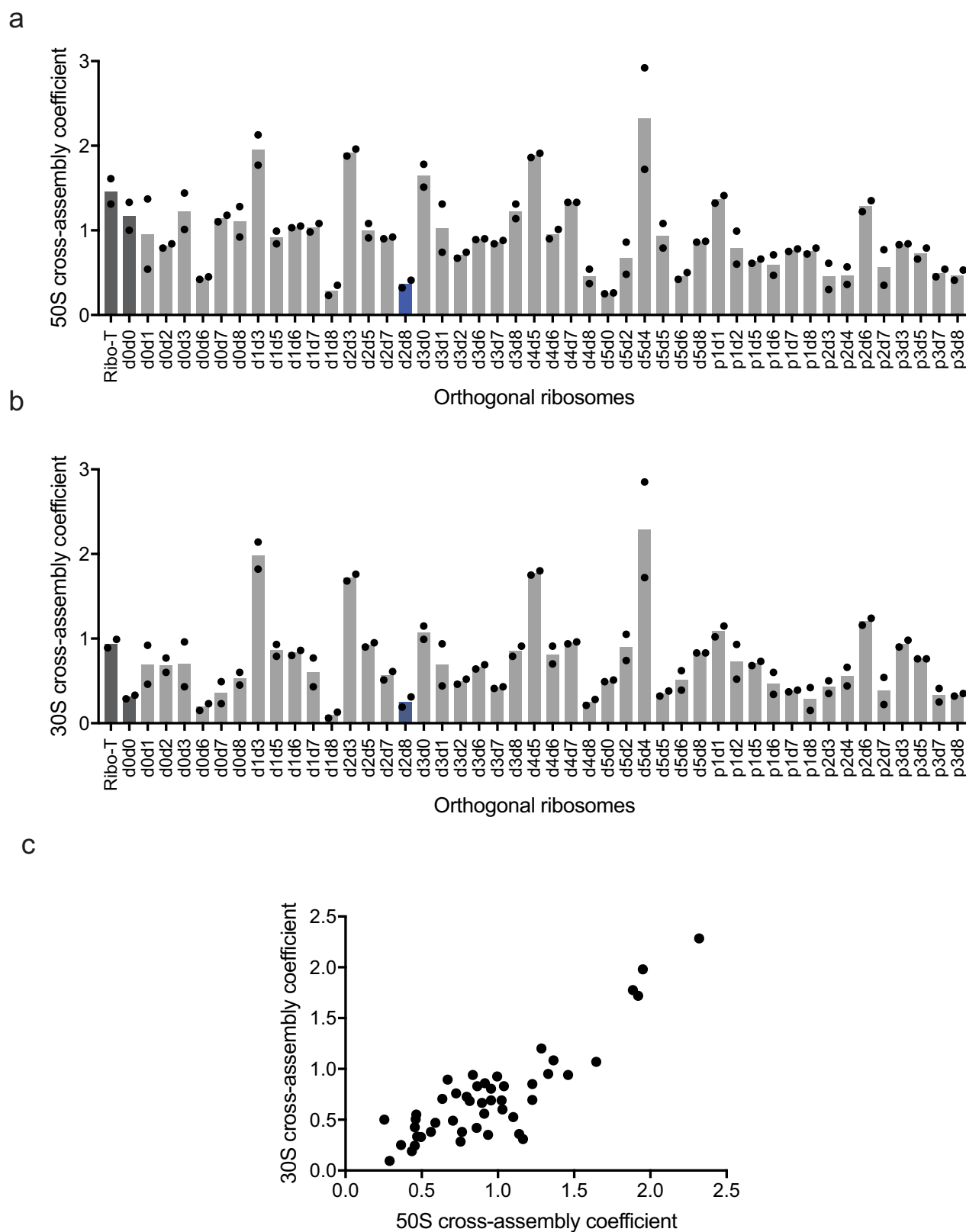
b, In vivo activity of linker-length variants of O-stapled ribosomes (using the same dataset as in Fig. 2b). GFP expression was analysed in *E. coli* cells containing the indicated O-ribosome, the *Methanosarcina mazei* PylRS synthetase/tRNA_{CUA} pair, and the O-sfGFP150_{TAG} reporter, in the presence of 1 mM Bock (N-epsilon (tert-butoxycarbonyl)-L-lysine). GFP fluorescence is shown as a percentage of that produced from an orthogonal ribosome with independent, non-linked subunits. O-d2d8 is highlighted in blue, and O-ribosomes with previously described subunit linkers are in dark grey. Statistics are detailed in the Methods.



Extended Data Fig. 2 | See next page for caption.

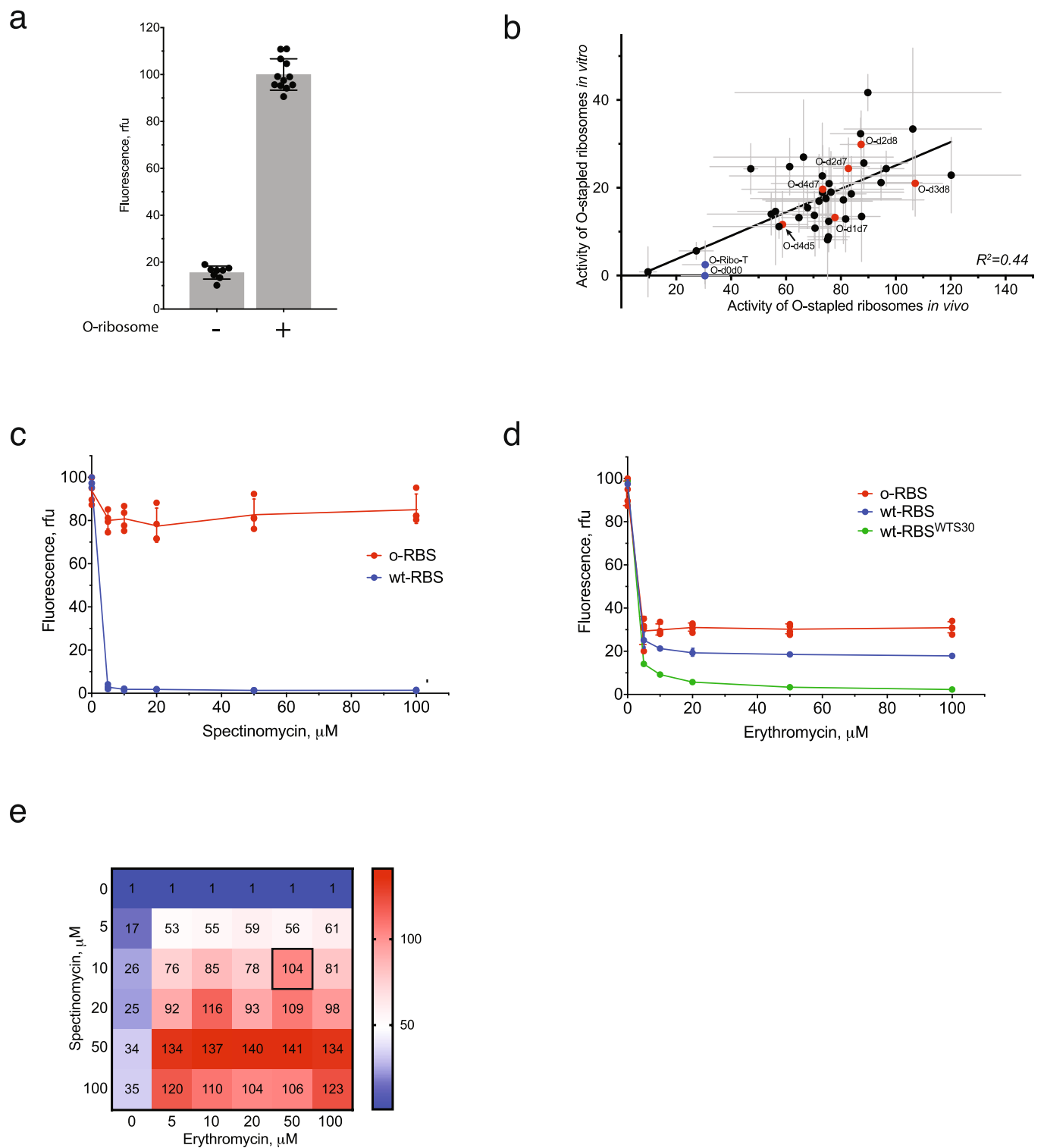
Extended Data Fig. 2 | MS2-tagged O-stapled ribosomes. **a**, Tagging O-stapled ribosomes with an MS2 stem loop minimally perturbs in vivo ribosome activity. We measured in vivo ribosome activities via GFP production from an O-sfGFP150_{TAG} reporter, in cells expressing an intact or MS2-tagged linker-length variant of O-stapled ribosome along with the *M. mazei* pyrrolysyl-tRNA synthetase/tRNA_{CUA} pair (encoded by *PylS/pylT*) in the presence of 1 mM BocK. GFP fluorescence was normalized to that produced from a non-stapled O-ribosome. For numbers of replicates (*n*) and statistics, see Methods and Supplementary Data 2. **b**, Sucrose gradient analysis of an *E. coli* lysate with and without an O-stapled ribosome variant; *n* = 3 biological replicates. **c**, Affinity purification of a non-stapled O-ribosome depends on the presence of GST-tdMS2CP (a fusion between glutathione-S-transferase (GST) and

a mutant of a tandem dimer of the MS2 coat protein (tdMS2CP)) on resin. Affinity purification of MS2-tagged ribosomes was performed on glutathione-sepharose resin with bound GST-tdMS2CP (lanes 2–4) and without GST-tdMS2CP (lane 1). Varying amounts of total RNA were loaded in lanes 2–4. **d**, Affinity purification depends on the presence of the MS2 RNA stem loop on ribosomes. Pellets of O-p2d6 ribosomes were collected through sucrose cushions. Affinity purifications were performed on glutathione-sepharose resin with bound GST-tdMS2CP. **e**, O-d5d8-MS2 was affinity purified, and MS2 stem-loop-containing species were probed by northern blot (NB). EtBr, ethidium bromide (a fluorescent stain for nucleic acid). The experiments in **c–e** were each performed once. For source data regarding gels, see Supplementary Fig. 1.



Extended Data Fig. 3 | Engineered O-stapled ribosome variants minimize cross-assembly. **a**, Screen of 50S cross-assembly coefficients for different O-ribosomes with linked subunits. $n = 2$ biological replicates; each replicate is shown by a dot, and the bars represent the means (using the same dataset as in Fig. 2d). **b**, Screen of 30S cross-assembly coefficients. $n = 2$ biological replicates; each replicate is shown by a dot, and the bars represent the means (using the same dataset as in Fig. 2e).

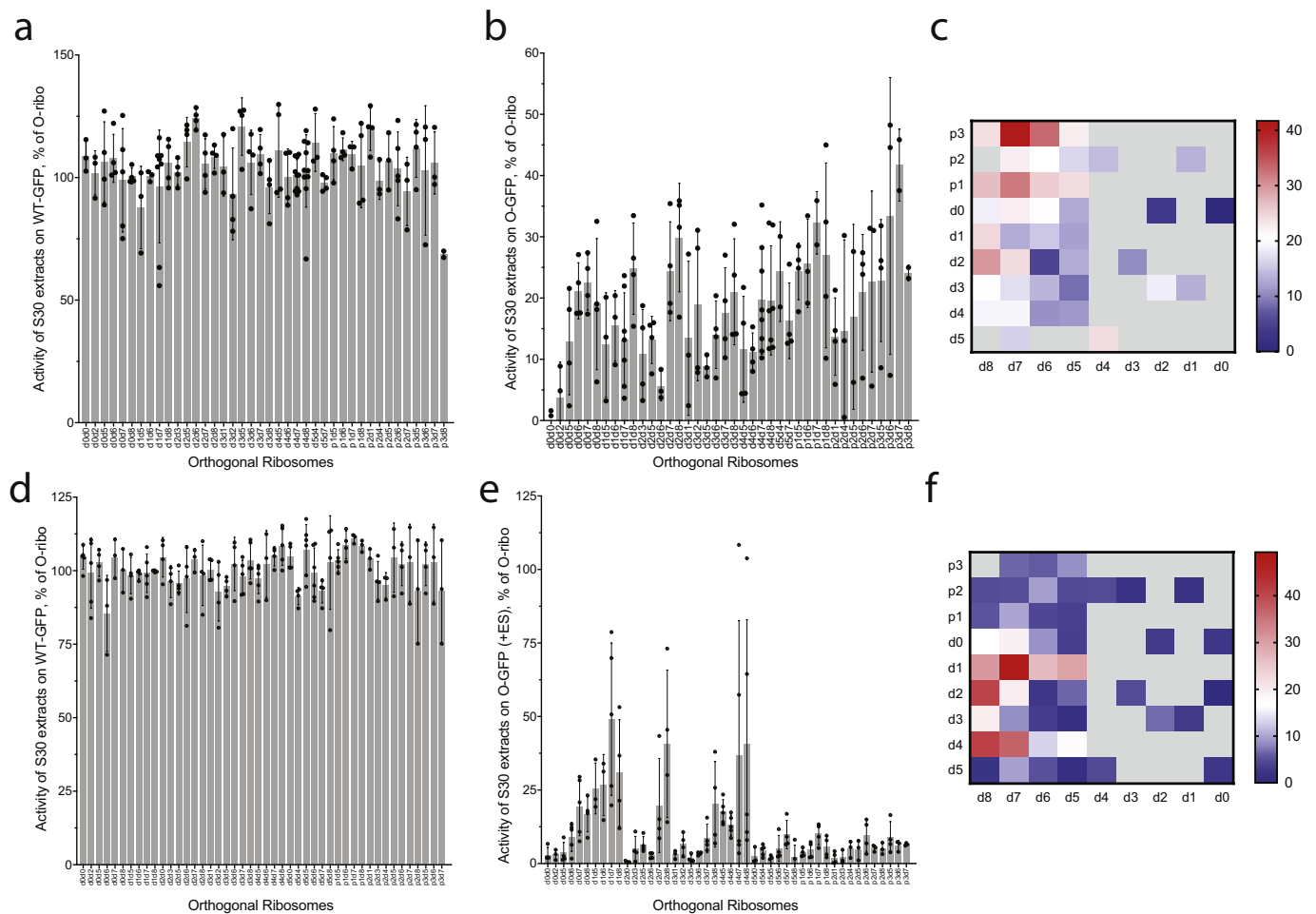
In **a** and **b**, more than 90% of ribosomes had cross-assembly coefficients between 0 and 1, as expected. Previously reported O-ribosomes with linked subunits are shown in dark grey, while O-d2d8 is highlighted in blue. **c**, Correlation between the means (from $n = 2$ biological replicates) of 50S and 30S cross-assembly coefficients for different O-stapled ribosome variants; the variation in these data is shown in **a**, **b**.



Extended Data Fig. 4 | See next page for caption.

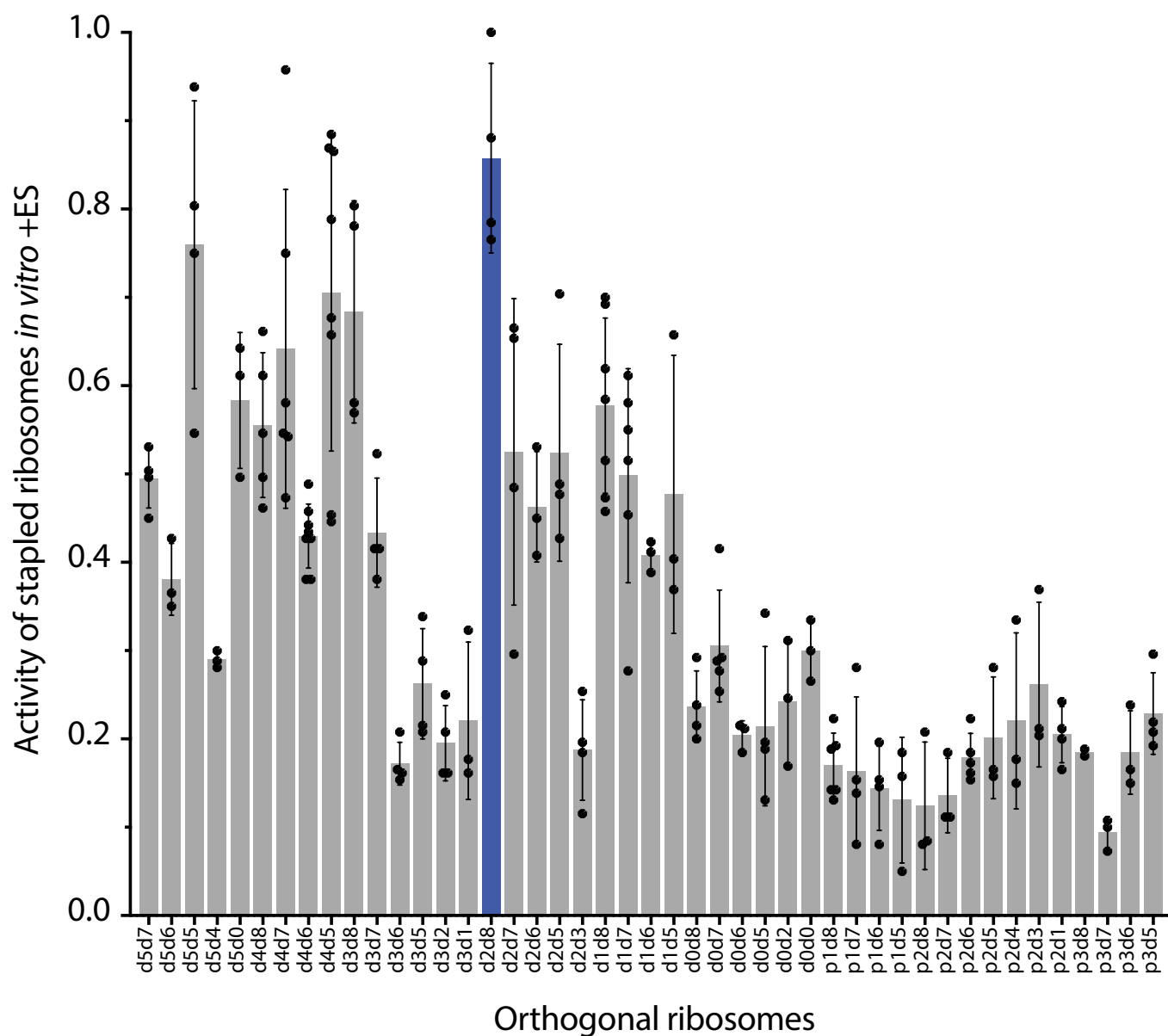
Extended Data Fig. 4 | In vitro translation activity of O-stapled ribosomes upon inhibition of native ribosomes. **a**, In vitro translation of T7-O-GFP in S30 extracts with and without the non-stapled O-ribosome. The data show the means (grey bars) of, respectively, 12 and 8 independent replicates (dots); the error bars show \pm s.d. **b**, In vitro translation activities of O-stapled ribosome variants (y -axis) were measured via the GFP fluorescence produced from T7-O-GFP. In vivo activities (measured as described in Extended Data Fig. 1b) are shown on the x -axis. Individual replicates are shown in Extended Data Fig. 1b, and statistics are described in the Methods. **c**, In vitro translation of T7-O-GFP (O-RBS) or T7-GFP (wt-RBS) in S30 extracts containing the non-stapled O-ribosome in the presence of spectinomycin. The O-16S rRNA of the O-ribosome contains the C1192U mutation, which confers resistance to spectinomycin. The data show the mean of $n = 4$ independent replicates; error bars are \pm s.d. From this, we conclude that 10 μ M of spectinomycin is sufficient to inhibit the translational activity of wild-type small subunits in the S30 extract, but has minimal effect on spectinomycin-resistant subunits. rfu, relative

fluorescence units. **d**, In vitro translation of T7-O-GFP (O-RBS) or T7-GFP (wt-RBS) in S30 extracts containing the non-stapled O-ribosome in the presence of erythromycin. The 23S rRNA that is co-expressed with the O-16S rRNA contains the A2058G mutation, which confers resistance to erythromycin. In vitro translation of T7-GFP (wt-RBS^{WTS30}) was also performed in S30 extracts without the O-ribosome. The data show the means of $n = 4$ independent replicates; error bars show \pm s.d. We found that 50 μ M erythromycin reduces translation from the wild-type ribosome-binding site to 18.5% of the level seen without erythromycin, and reduces translation from the orthogonal ribosome-binding site to 30% of the level without erythromycin. We conclude that 50 μ M erythromycin is sufficient to inhibit wild-type large subunits in the S30 extract. **e**, Ratios of GFP produced from T7-O-GFP versus T7-GFP in S30 extracts containing the non-stapled O-ribosome are shown as a function of spectinomycin and erythromycin concentration. The data are means of $n = 4$ independent replicates; statistics are in the Methods.



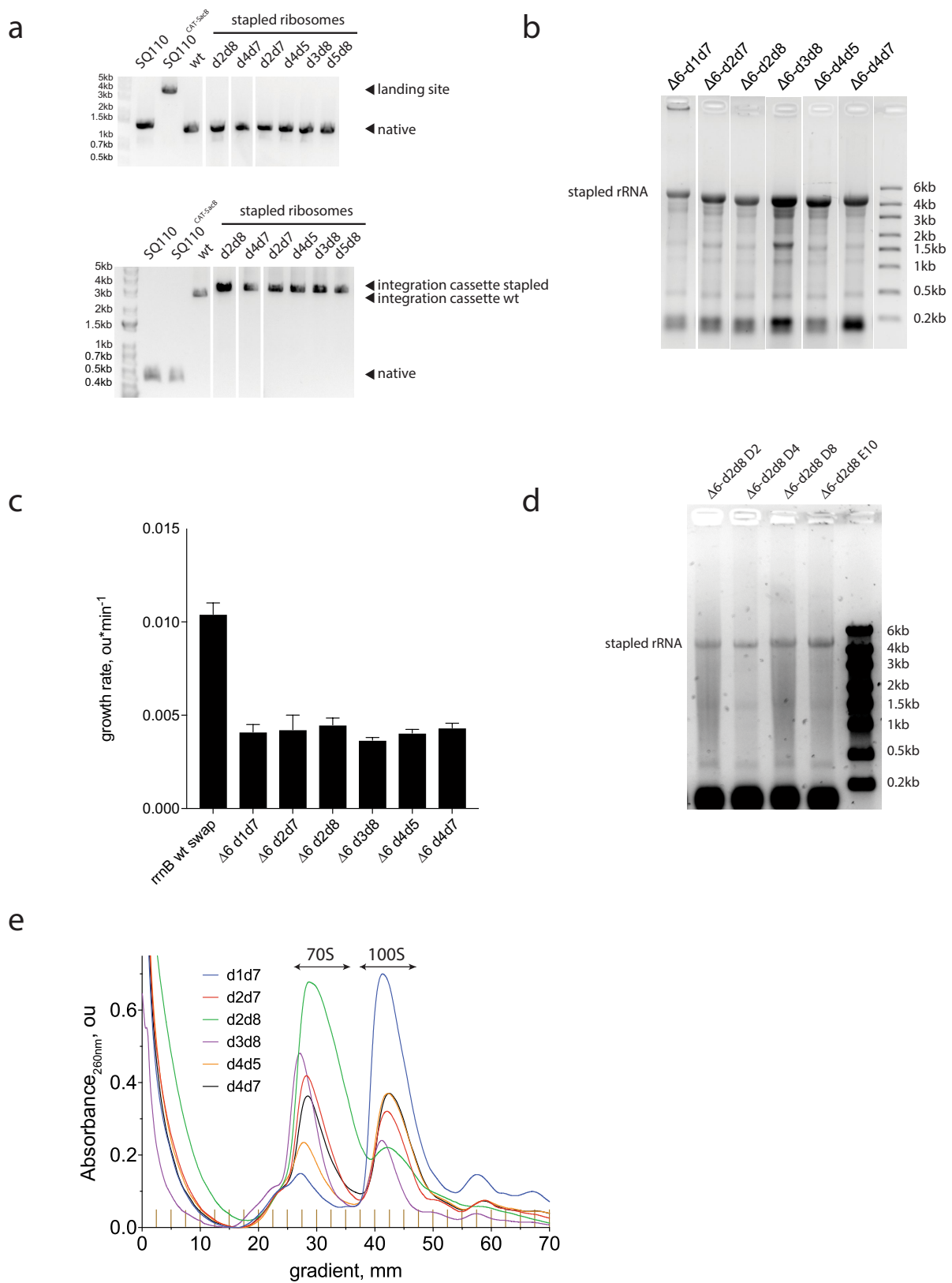
Extended Data Fig. 5 | Activities of O-stapled-ribosome linker-length variants in cell-free protein synthesis. **a**, Activities of lysates that contain a mixture of host and O-stapled ribosomes on a WT-GFP reporter, showing that the translation machinery in the tested lysates is equally active. Every data point shows the activity of an independently produced S30 extract from independently grown cells. **b**, Activities of the same set of lysates as in **a** on an O-GFP reporter, normalized to their activity on the WT-GFP reporter. Every data point shows the activity of an independently produced S30 extract from independently grown cells. Detailed statistics are shown in Supplementary Data 3. **c**, The same dataset as in **b**, represented as a heatmap for ease of comparison. **d**, Activity of lysates containing a mixture of host and O-stapled ribosomes on the

WT-GFP reporter. Every data point shows independently measured activities of O-stapled ribosome variants in independently prepared S30 extracts. **e**, Activity of the same set of the lysates as in **d** on the O-GFP reporter in the presence of 10 μM spectinomycin and 50 μM erythromycin, normalized to their activity on WT-GFP. Data points are independently measured activities of O-stapled ribosome variants in independently prepared S30 extracts. Detailed statistics are given in Supplementary Data 4. **f**, The same dataset as in **e**, represented as a heatmap for ease of comparison. In **a**, **b**, **d** and **e**, the error bars show \pm s.d., the grey bar shows the mean of the number (n) of independent replicates. Values for n are given in the Methods.



Extended Data Fig. 6 | In vitro translation activity of O-stapled ribosome variants upon inhibition of native ribosome subunits. This figure uses the same dataset as in Fig. 2g, and shows the *in vitro* translation of T7-O-GFP in S30 extracts containing different O-stapled ribosome variants in the presence of 10 μ M spectinomycin and 50 μ M erythromycin

(which inhibit contributions to translation from endogenous subunits). The dots indicate individual data for each tested O-stapled ribosome, and the grey bars show means, for the number of independent replicates given in the Methods. O-d2d8 is highlighted in blue, and the error bars indicate \pm s.d.

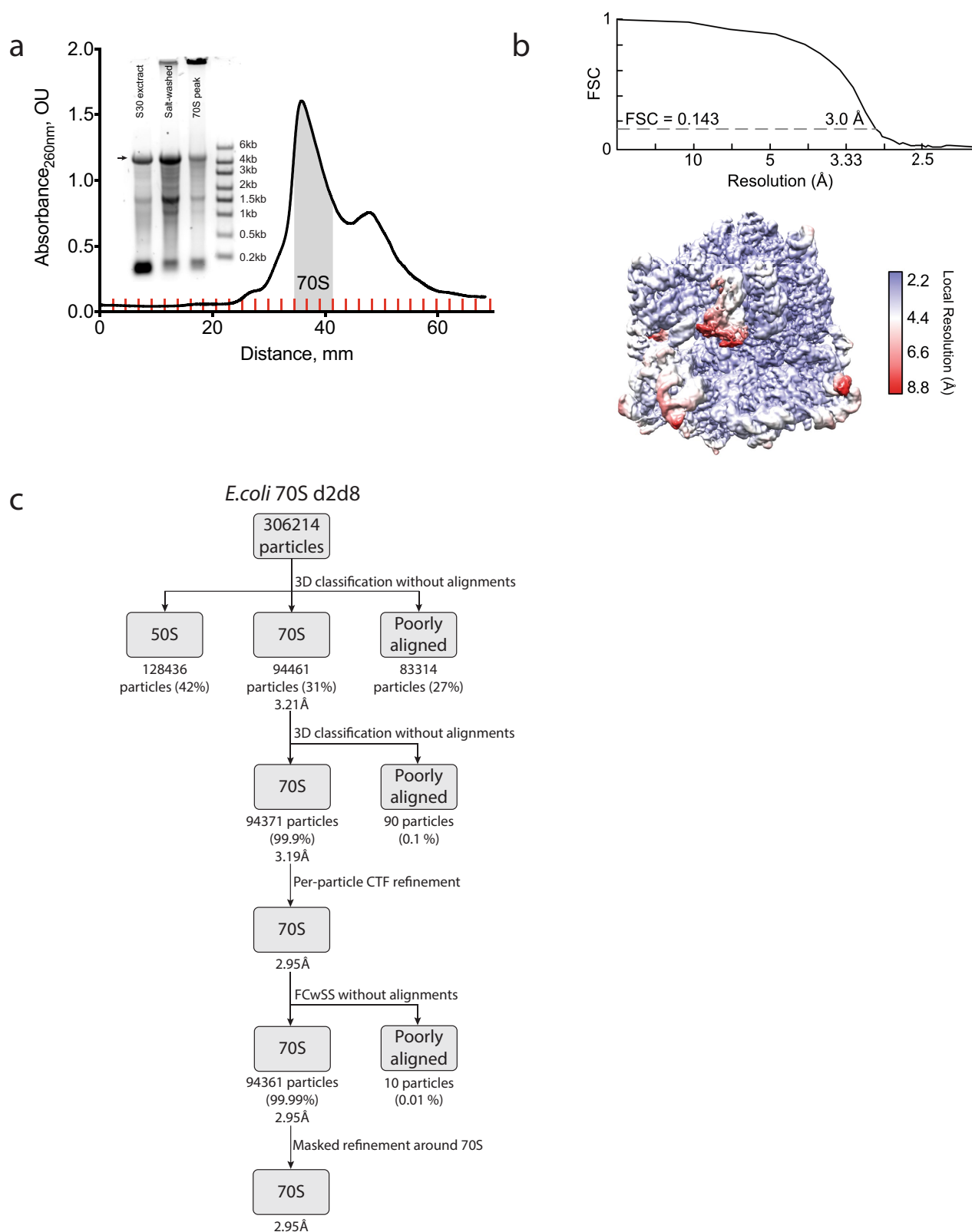


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Genomically encoded stapled ribosomes.

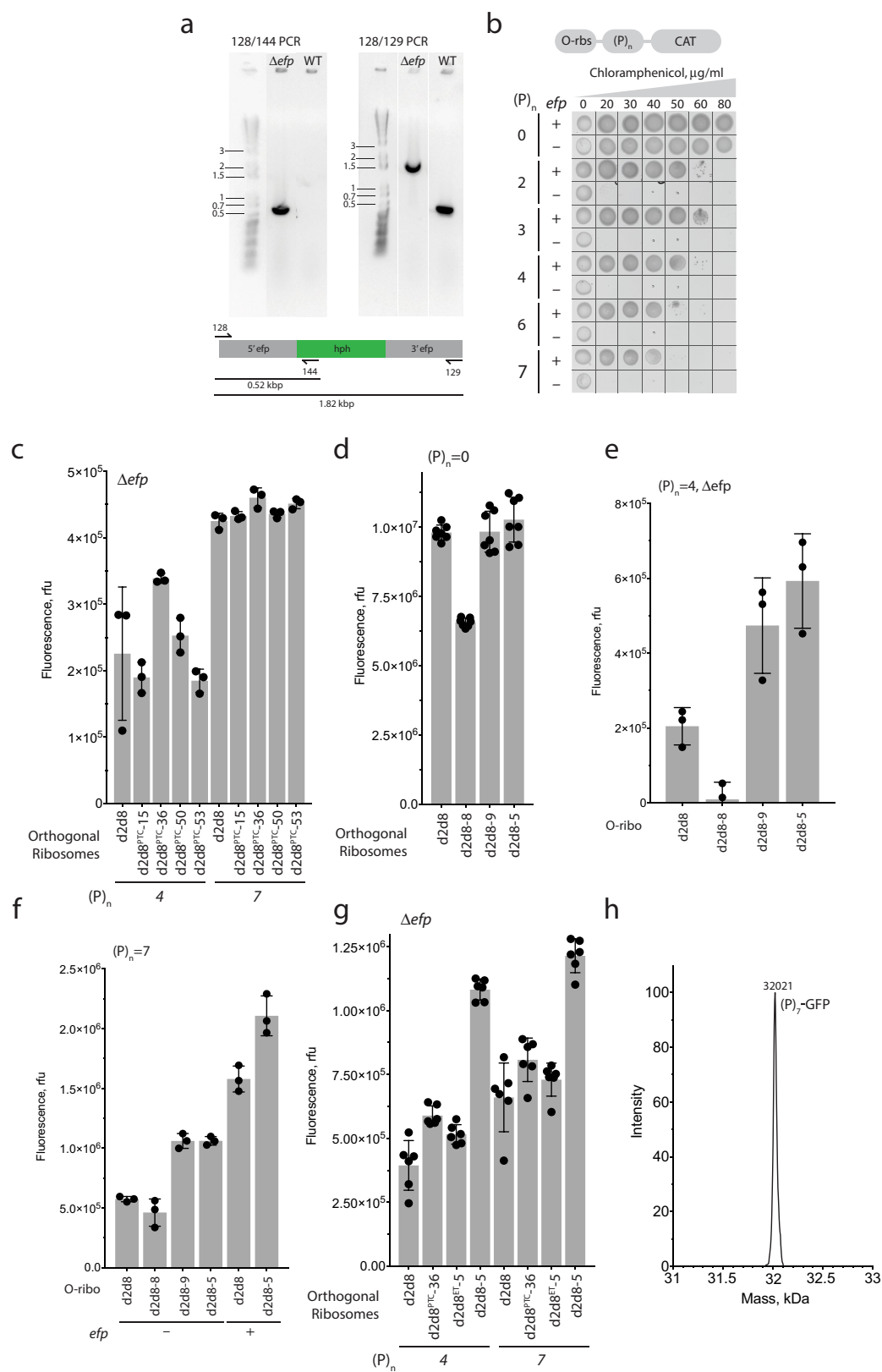
a, Colony PCR products reflect the genomic exchanges seen in *E. coli* strains with a genomically encoded, stapled ribosome rRNA operon integrated by ribo-REXER. Top, direct amplification of the genomic region upstream of the *rrnE* locus. Integration of the landing site (SacB/CAT) to create the strain SQ110^{CAT-SacB} increases the length of the region upstream of *rrnE* from 1.3 kb to 3.5 kb. After ribo-REXER, this cassette is lost again, leading to a 1.3-kb band. Bottom, amplification from nucleotide 2,630 of 23S to the *rrnE* terminator region shows integration of the PheS*/HygR cassette after integration of the wild-type *rrnB* operon, indicated by 'integration cassette wt'. Integration of a stapled-ribosome cassette increases the length of the PCR product further, as it includes the 16S 3' and internal transcribed spacer (ITS) regions, leading to the bands

indicated by 'integration cassette stapled'. The experiment was performed once. **b**, Denaturing RNA gel electrophoresis reveals the expression of rRNA (around 4,500 nucleotides) from intact stapled ribosomes as the predominant RNA species in all strains. The experiment was performed once. **c**, Growth rates of strains after successful ribo-REXER. WT describes the integration of a non-stapled *rrnB* operon. For statistics, see Methods. **d**, Denaturing RNA gel electrophoresis shows the expression of rRNA from intact stapled ribosomes (around 4,500 nucleotides) as the predominant RNA species in all evolved $\Delta 6$ d2d8 strains. The experiment was performed once. **e**, Sucrose gradient analyses of stapled ribosomes isolated from cells under ribosome-associating conditions (10 mM MgCl₂); the experiments were repeated three times with similar results. For source data regarding gels, see Supplementary Fig. 1.



Extended Data Fig. 8 | Purification of the d2d8 stapled ribosome and structural determination by cryo-EM. **a**, Isolation of the d2d8 stapled ribosome on a sucrose gradient for cryo-EM. Fractions corresponding to the middle of the 70S peak (grey shading) were collected and used in structural studies. The inset shows RNA agarose gel analysis of the stapled ribosome at the different purification stages (30S extract preparation and salt wash) as well as the combined 70S fraction sample. The arrow indicates the position of the stapled rRNA. For source data regarding

gels, see Supplementary Fig. 1. The data represent $n = 2$ independent preparations. **b**, Fourier shell correlation (FSC) curve, calculated between independent half-maps. The resolution is estimated from the map-to-map correlation at FSC = 0.143 (for a detailed description, see Extended Data Table 1). The electron-microscopy map is coloured according to local resolution. **c**, Workflow showing the three-dimensional classification and refinement of cryo-EM particles.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Selection of O-ribosome variants competent in translation of polypyrroline sequences. **a**, PCR characterization of an *E. coli* TOP10 Δefp strain. PCR was carried out, using the indicated primer pairs (128/144 or 128/129), on genomic DNA from a TOP10 strain with intact *efp*, or a strain in which the *efp* locus was disrupted by a hygromycin-B-resistance gene (*hph*). For source data regarding gels, see Supplementary Fig. 1. The experiment was performed twice. **b**, Cell-growth assay, showing the translation activity of O-d2d8 on O-(P)_n-CAT reporters in wild-type (+) or Δefp (–) *E. coli* TOP10 cells and with varying concentrations of chloramphenicol. The experiment was performed once. **c**, Activity of the PTC-library hits and the parental O-d2d8 ribosome on p15A-O-(P)₄-GFP and p15A-O-(P)₇-GFP reporters in *E. coli* TOP10 Δefp cells. $n = 3$ biological replicates; error bars indicate \pm s.d. **d**, Translation activity of exit-tunnel library hits in the absence of proline-rich sequences. p15A-O-

GFP was used as a reporter. $n = 7$ biological replicates; error bars indicate \pm s.d. **e**, GFP fluorescence resulting from translation of O-(P)₄-GFP in *E. coli* TOP10 Δefp cells. $n = 3$ biological replicates; error bars indicate \pm s.d. **f**, As for **e**, except that the activity of the evolved mutants was tested on an O-(P)₇-GFP reporter in *E. coli* with and without *efp*. $n = 3$ biological replicates; error bars represent \pm s.d. **g**, Mutations in both the PTC and the exit tunnel are required to confer on O-d2d8 the ability to translate O-(P)₄-GFP and O-(P)₇-GFP in TOP10 Δefp cells. O-d2d8^{PTC}-36 and O-d2d8^{ET}-5 contain selected mutations in the PTC and the exit tunnel respectively; O-d2d8-5 contains mutations in both. See Supplementary Data 12 for details on mutations. $n = 6$ biological replicates; error bars represent \pm s.d. **h**, Electrospray ionization spectra of O-(P)₇-GFP synthesized by O-d2d8-5 in Δefp *E. coli*. This experiment was performed once.

Extended Data Table 1 | Collection, refinement and validation of cryo-electron-microscopy data

| | |
|--|--|
| | 70S d2d8 stapled (EMDB-0261) (PDB-6HRM) |
| Data collection and processing | |
| Magnification | 134615x |
| Voltage (kV) | 300 |
| Electron exposure (e-/Å ²) | 26.85 |
| Defocus range (μm) | -2.0 to -3.5 |
| Pixel size (Å) | 1.06 |
| Symmetry imposed | none |
| Initial particle images (no.) | 306214 |
| Final particle images (no.) | 94371 |
| Map resolution (Å) | 3.0 |
| FSC threshold | 0.143 |
| Map resolution range (Å) | 2.2-8.8 |
| Refinement | |
| Initial model used (PDB code) | 5MDZ |
| Model resolution (Å) | 3.1 |
| FSC threshold | 0.143 |
| Model resolution range (Å) | 2.6-7.1 |
| Map sharpening <i>B</i> factor (Å ²) | -78.8 |
| Model composition | |
| Non-hydrogen atoms | 145181 |
| Protein residues | 6131 |
| Nucleic acid residues | 4563 |
| Ligands | 453 |
| <i>B</i> factors (Å ²) | |
| Protein | 86.4 |
| RNA | 55.1 |
| R.m.s. deviations | |
| Bond lengths (Å) | 0.008 |
| Bond angles (°) | 1.004 |
| Validation | |
| MolProbity score | 1.45 |
| Clashscore | 2.96 |
| Poor rotamers (%) | 0.47 |
| Ramachandran plot | |
| Favored (%) | 94.78 |
| Allowed (%) | 5.16 |
| Disallowed (%) | 0.05 |

Elasticity of lower-mantle bridgmanite

ARISING FROM A. Kurnosov, H. Marquardt, D. Frost, T. Boffa Ballaran & L. Ziberna *Nature* **543**, 543–546 (2017); <https://doi.org/10.1038/nature21390>

Bridgmanite ((Mg,Fe)(Fe,Al,Si)O₃) is believed to be the most abundant mineral in Earth's lower mantle, accounting for approximately 75% of the region by volume^{1,2}. Recently, Kurnosov et al.³ reported on the single-crystal elasticity of (Al,Fe)-bearing bridgmanite up to 40 GPa and concluded that a Fe³⁺/Fe²⁺ ratio of about two is required to match Preliminary Reference Earth Model (PREM) seismic profiles⁴ at depths below 1,200 km. Here we use a sensitivity test between measured velocities and elastic constants (C_{ij}) and a covariance matrix analysis to show that their derived elastic constants and velocities have large uncertainties at lower-mantle pressures. These uncertainties invalidate the assertion of ref.³ of the existence of an Fe³⁺-rich pyroclitic lower mantle with (Al,Fe)-bearing bridgmanite, and therefore further experimental studies at relevant lower-mantle pressure–temperature conditions are required to reliably infer the mineralogical and geochemical properties of the deep mantle. There is a Reply to this Comment by Kurnosov, A. et al. *Nature* **564**, <https://doi.org/10.1038/s41586-018-0742-6> (2018).

Bridgmanite contains approximately 10% iron (in Fe²⁺ and Fe³⁺ states) and 5%–7% Al₂O₃ in its structure. Knowledge of the elasticity of (Al,Fe)-bearing bridgmanite at relevant pressure–temperature conditions is thus essential for comparison with the seismic models used to improve our understanding of the physics and chemistry of the lower mantle. In recent decades, there have been extensive experimental

and theoretical efforts to obtain single-crystal elastic constants and polycrystalline aggregate compressional and shear-wave velocities (V_P and V_S) for (Al,Fe)-bearing bridgmanite at high-pressure and high-temperature conditions^{5–11}.

Bridgmanite crystallizes in an orthorhombic crystal structure at lower-mantle pressure–temperature conditions and exhibits nine independent C_{ij} . The longitudinal moduli (C_{11} , C_{22} and C_{33}) strongly correlate with the aggregate V_P value, whereas the shear moduli (C_{44} , C_{55} and C_{66}) determine the aggregate V_S value (Fig. 1, Supplementary Information). Reliable derivation of the nine C_{ij} in bridgmanite requires extensive experimental measurements of both V_P and V_S for crystallographic orientations that are intrinsically sensitive to longitudinal, shear and off-diagonal C_{ij} (Fig. 1a, b; Extended Data Fig. 1). Kurnosov et al.³ used Brillouin scattering to measure the V_P and V_S values of single-crystal (Al,Fe)-bearing bridgmanite (Mg_{0.9}Fe_{0.1}Si_{0.9}Al_{0.1}O₃) and X-ray diffraction to determine the crystallographic orientations and densities of two platelets in a high-pressure diamond anvil cell up to approximately 40 GPa. Analysis of the sensitivity of the crystal orientations between each C_{ij} , the direction-dependent velocities (V_P , V_{S1} and V_{S2}), and the number of measured phonon directions show that the crystal orientations in ref.³ are appropriate for tightly constraining only some of the shear moduli¹²: C_{44} and C_{55} can be determined with

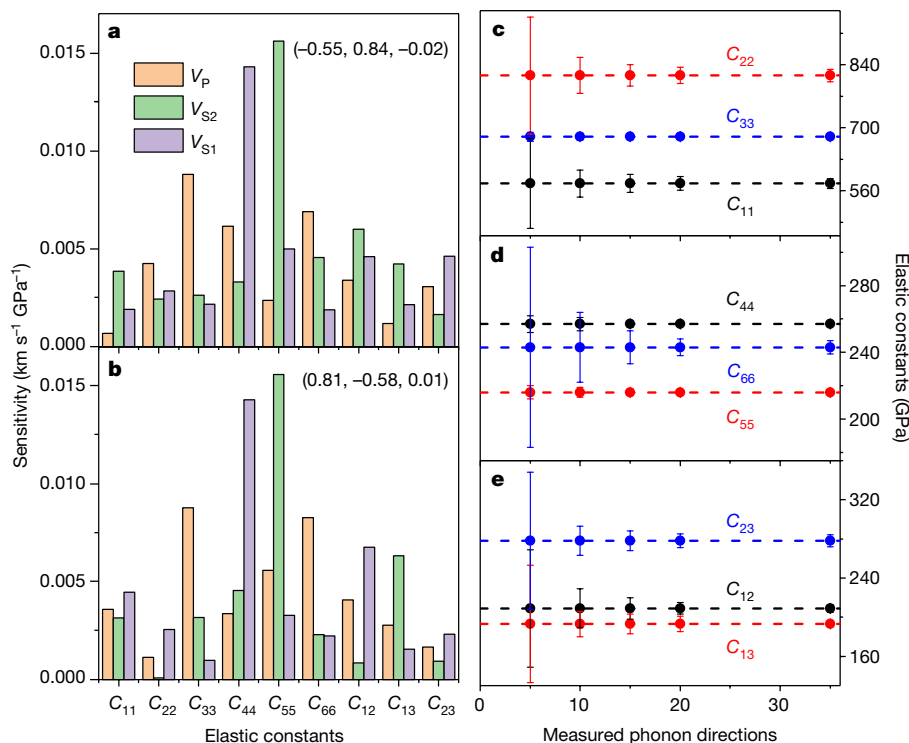


Fig. 1 | C_{ij} sensitivity to the velocities and number of measured phonon directions of single-crystal bridgmanite in two crystallographic orientations used in ref.³ at 40.17 GPa. a, b, The sensitivity of V_P and two orthogonally polarized shear-wave velocities V_{S1} and V_{S2} in two orientations. The orientations of the two crystal platelets are given in

parentheses. c–e, The sensitivity is used to evaluate the uncertainties of the elastic constants as a function of the number of measured phonon directions with 10° azimuthal separation (see Supplementary Information for details). Error bars represent standard deviations (1 σ) of the elastic moduli.

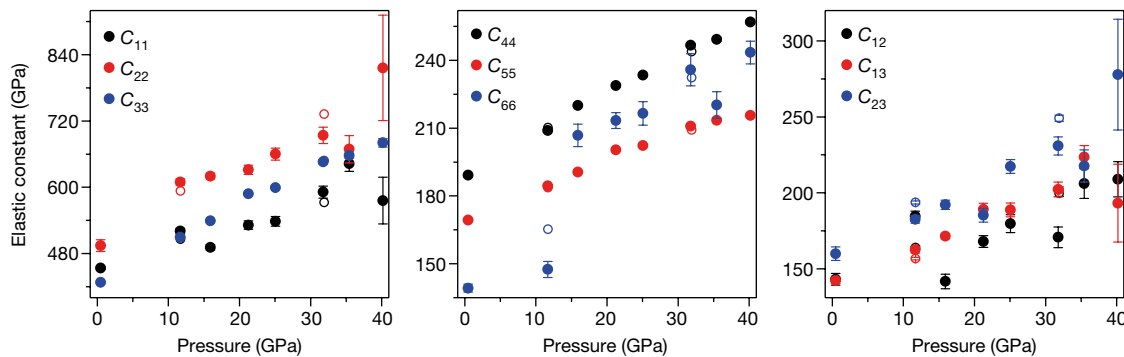


Fig. 2 | Derived full elastic constants of (Al,Fe)-bearing bridgmanite at high pressure. Filled data points are our calculated results using the raw velocity data in ref. ³ and the covariance matrix analysis; open data points are updated results in the Author Correction to ref. ³ using the global fit

method (only two pressure points were provided in the first draft of their Reply). Error bars represent standard deviations (1σ) and are not shown when smaller than the symbols.

high sensitivity, but the orientations examined provide only weak constraints on longitudinal and off-diagonal moduli (Fig. 1c–e). We note that measuring the V_P of bridgmanite using Brillouin spectroscopy at high pressures in a diamond anvil cell is technically difficult because the sample's V_P is mostly blocked by the diamond V_S peak at pressures above approximately 23 GPa—the starting pressure at the top of the lower mantle. At 40.17 GPa, Kurnosov et al. ³ measured 15 phonon directions in V_P and 30 phonon directions in V_S . Based on our modelling, the uncertainties of the longitudinal and shear moduli at 40.17 GPa are approximately 2% and 0.5%–1%, respectively (Fig. 1c, d).

Christoffel's equation exactly defines the relationship between C_{ij} , the direction-dependent velocities, and the density of bridgmanite, and has been used to derive C_{ij} from velocity data. Finite-strain equations can also be applied to model high-pressure results to investigate the effects of pressure on elasticity. However, because there are nine C_{ij} variables involved in the analysis, the covariance matrix method is necessary to rigorously evaluate the uncertainties and correlations of the derived C_{ij} (Fig. 2) ¹³. Our analysis of the data from ref. ³ using this method clearly shows that most C_{ij} above 25 GPa exhibit strong trade-offs of modelled values and their uncertainties among one another, with increased

uncertainties at higher pressures (Supplementary Tables 2–9). In light of this, we consider the errors reported for C_{ij} in Kurnosov et al. ³ to be unrealistically low (Fig. 2). Kurnosov et al. ³ used a 'global fit' method to simultaneously account for all measured direction-dependent acoustic-wave velocities and densities at all experimental pressures. However, analysis of the global fit method and the derived C_{ij} by Kurnosov et al. ³ shows that the errors of C_{ij} at higher pressures above 25 GPa (with limited V_P data) are drastically reduced by lower-pressure data that have more V_P and V_S measurements with smaller errors. Since Christoffel's equation exactly defines the relationship between C_{ij} and the direction-dependent velocities, the greater errors in C_{ij} at higher pressures cannot be simply reduced and compensated by results obtained at lower pressures, even when the finite-strain equations are used concurrently in the modelling. Furthermore, analysis of the reported C_{ij} values and experimental velocity data in Kurnosov et al. ³ using Christoffel's equation at two representative pressures of 11.66 GPa and 31.76 GPa shows that their velocity data are inconsistent with the reported C_{ij} : the velocity fitting curve cannot be reproduced using their reported C_{ij} values, densities and crystallographic orientations (Extended Data Figs. 2 and 3).

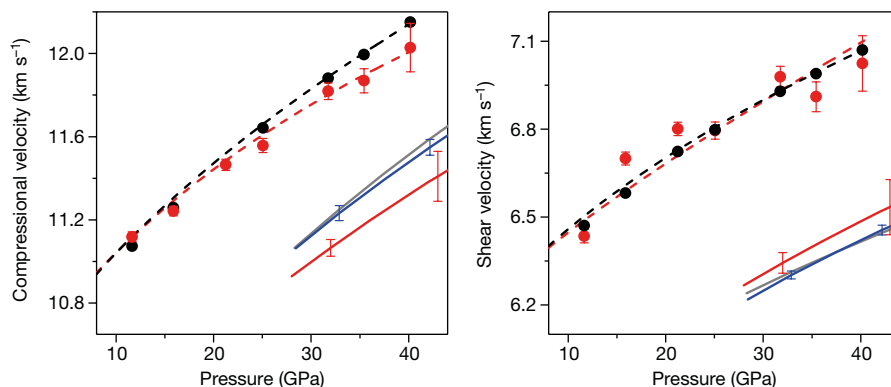


Fig. 3 | V_P and V_S profiles of the (Al,Fe)-bearing bridgmanite ($\text{Mg}_{0.9}\text{Fe}_{0.1}\text{Si}_{0.9}\text{Al}_{0.1}\text{O}_3$) in the lower mantle. V_P and V_S profiles in PREM ⁴ are also plotted for comparison (grey lines). Red filled circles are our calculated V_P and V_S values of the bridgmanite at high pressures and 300 K using our obtained adiabatic bulk modulus (K_S) and shear modulus (G); the red dashed lines are the finite-strain equation fits to our calculated data; the black filled circles are the V_P and V_S values of the bridgmanite at high pressures and 300 K reported in figure 2 of ref. ³; and the black dashed lines are the finite-strain equation fits to their results. The solid red lines are our newly calculated aggregate V_P and V_S values for a simplified pyrolite model containing 80 vol% bridgmanite ($(\text{Mg}_{0.9}\text{Fe}_{0.1}\text{Si}_{0.9}\text{Al}_{0.1})$

O_3) and 20 vol% ferropericlase ($(\text{Mg}_{0.8}\text{Fe}_{0.2})\text{O}$) along a representative geotherm using thermoelastic parameters reported by refs. ³ and ¹⁵. The solid blue lines represent V_P and V_S values of a pyrolitic mantle along a representative geotherm reported by ref. ³. Error bars represent standard deviations (1σ). The standard deviations of the red lines are estimated from our results using the sensitivity and covariance matrix analysis with standard error propagation, while the uncertainties of the blue lines are from ref. ³. We note that our calculated V_P and V_S along a representative geotherm exhibit large uncertainties and could not be used to reliably constrain the lower-mantle composition.

Using the derived C_{ij} , the aggregate V_p and V_s of the bridgmanite can be calculated at high pressure and 300 K (Fig. 3; Extended Data Figs. 4–6). (Al,Fe)-bearing bridgmanite is only stable above approximately 23 GPa, and our analysis shows the uncertainties of both V_p and V_s to be about 0.5% at this pressure. At the highest pressure reported by ref. ³ (40.17 GPa), they report uncertainties in V_p and V_s of approximately 1.5% and 1.0%, respectively. The lower mantle is probably made up of 75% (Al,Fe)-bearing bridgmanite, 20% ferropericlase (approximately $(\text{Mg}_{0.8}\text{Fe}_{0.2})\text{O}$), and 5% calcium perovskite. Previous studies have also found that bridgmanite contains a significant amount of Fe^{3+} and that the partition coefficient of iron between bridgmanite and ferropericlase varies with increasing depth¹. To make the modelling more difficult, a small amount of metallic iron could be formed in the lower mantle because of the charge disproportionation reaction in bridgmanite¹⁴. Elastic parameters and their uncertainties for these mineral phases as well as element partitioning need to be considered at lower-mantle pressure–temperature conditions to realistically investigate the geophysical consequences of the contributions of these components to seismic profiles. Our analysis shows that the uncertainties of V_p and V_s of bridgmanite at 40 GPa along an expected geotherm (or around 1,000 km in depth) are approximately 2.0% and 1%, respectively. Although Kurnosov et al.³ stated that the V_p and V_s profiles of a pyrolitic mineralogical model are consistent with PREM⁴, their uncertainties are of the order of a few per cent when taking the aforementioned contributions into account. These uncertainties are larger than the magnitude of the metallic iron and Fe^{3+} content effects (less than 0.5% according to figure 4a in ref. ³) on the modelled lower-mantle V_p and V_s profiles. In fact, with such large uncertainties, the pyrolitic compositional model of the lower mantle can be called into question. Kurnosov et al.³ also implied the presence of metallic iron, a change in bridgmanite Al-Fe cation ordering, or a decrease in the ferric iron content in the lower parts of the lower mantle. However, at the current resolution of the elastic constant results of lower-mantle minerals, the existence of metallic iron and Fe^{3+} -rich bridgmanite would be seismically invisible.

In summary, we consider the use of a sensitivity test and covariance matrix to be a rigorous approach for the investigation of the elastic constants of single-crystal bridgmanite and its aggregate sound velocities at high pressure. At present, the uncertainties of the V_p and V_s profiles of (Al,Fe)-bearing bridgmanite at relevant lower-mantle conditions are too large to support the existence of Fe^{3+} -rich bridgmanite and a small amount of metallic iron in the lower mantle. Accordingly, further experimental studies to gather extensive velocity data of C_{ij} -sensitive orientations at relevant lower-mantle pressure–temperature conditions should be performed on candidate phases such as bridgmanite and ferropericlase to reliably infer the mineralogical and geochemical properties of the deep mantle.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Jung-Fu Lin^{1*}, Zhu Mao², Jing Yang^{1,3} & Suyu Fu¹

¹Department of Geological Sciences, Jackson School of Geosciences, The University of Texas at Austin, Austin, TX, USA. ²Laboratory of Seismology

and Physics of Earth's Interior, School of Earth and Planetary Sciences, University of Science and Technology of China, Hefei, China. ³Geophysical Laboratory, Carnegie Institution for Science, Washington, DC, USA.

*e-mail: afu@jsg.utexas.edu

Received: 7 February 2018; Accepted: 13 September 2018

Published online 19 December 2018.

1. Irfune, T. et al. Iron partitioning and density changes of pyrolite in Earth's lower mantle. *Science* **327**, 193–195 (2010).
2. Lin, J. F., Speziale, S., Mao, Z. & Marquardt, H. Effects of the electronic spin transitions of iron in lower mantle minerals: implications for deep mantle geophysics and geochemistry. *Rev. Geophys.* **51**, 244–275 (2013).
3. Kurnosov, A., Marquardt, H., Frost, D., Boffa Ballaran, T. & Ziberna, L. Evidence for a Fe^{3+} -rich pyrolitic lower mantle from (Al,Fe)-bearing bridgmanite elasticity data. *Nature* **543**, 543–546 (2017); Author Correction *Nature* **558**, E3 (2018).
4. Dziewonski, A. M. & Anderson, D. L. Preliminary Reference Earth Model. *Phys. Earth Planet. Inter.* **25**, 297–356 (1981).
5. Li, B. S. & Zhang, J. Z. Pressure and temperature dependence of elastic wave velocity of MgSiO_3 perovskite and the composition of the lower mantle. *Phys. Earth Planet. Inter.* **151**, 143–154 (2005).
6. Wentzcovitch, R., Karki, B., Cococcioni, M. & De Gironcoli, S. Thermoelastic properties of MgSiO_3 -perovskite: insights on the nature of the Earth's lower mantle. *Phys. Rev. Lett.* **92**, 018501 (2004).
7. Sinogeikin, S. V., Zhang, J. & Bass, J. D. Elasticity of single crystal and polycrystalline MgSiO_3 perovskite by Brillouin spectroscopy. *Geophys. Res. Lett.* **31**, L06620 (2004).
8. Sinogeikin, S. V., Sinogeikin, S. V., Hellwig, H., Bass, J. D. & Li, J. Sound velocity of MgSiO_3 perovskite to Mbar pressure. *Earth Planet. Sci. Lett.* **256**, 47–54 (2007).
9. Murakami, M., Ohishi, Y., Hirao, N. & Hirose, K. A perovskitic lower mantle inferred from high-pressure, high-temperature sound velocity data. *Nature* **485**, 90–94 (2012).
10. Jackson, J. M., Zhang, J., Shu, J., Sinogeikin, S. V. & Bass, J. D. High-pressure sound velocities and elasticity of aluminous MgSiO_3 perovskite to 45 GPa: implications for lateral heterogeneity in Earth's lower mantle. *Geophys. Res. Lett.* **32**, L21305 (2005).
11. Fu, S. et al. Abnormal elasticity of Fe-bearing bridgmanite in the Earth's lower mantle. *Geophys. Res. Lett.* **45**, 4725–4732 (2018).
12. Yoneda, A. et al. Elastic anisotropy of experimental analogues of perovskite and post-perovskite help to interpret D'' diversity. *Nat. Commun.* **5**, 3453 (2014).
13. Abramson, E., Brown, J., Slutsky, L. & Zaug, J. The elastic constants of San Carlos olivine to 17 GPa. *J. Geophys. Res. Solid Earth* **102**, 12253–12263 (1997).
14. Frost, D. J. et al. Experimental evidence for the existence of iron-rich metal in the Earth's lower mantle. *Nature* **428**, 409–412 (2004).
15. Stixrude, L. & Lithgow-Bertelloni, C. Thermodynamics of mantle minerals—I. Physical properties. *Geophys. J. Int.* **162**, 610–632 (2005).

Author contributions J.-F.L. initiated the project. Z.M. performed the data processing using the covariance matrix. J.Y. and S.F. performed the sensitivity test analysis. Z.M. and S.F. performed elasticity modelling and error analysis. S.F. prepared the draft Supplementary Information. J.-F.L. wrote the manuscript and all authors participated in the manuscript revision.

Competing interests Declared none.

Additional information

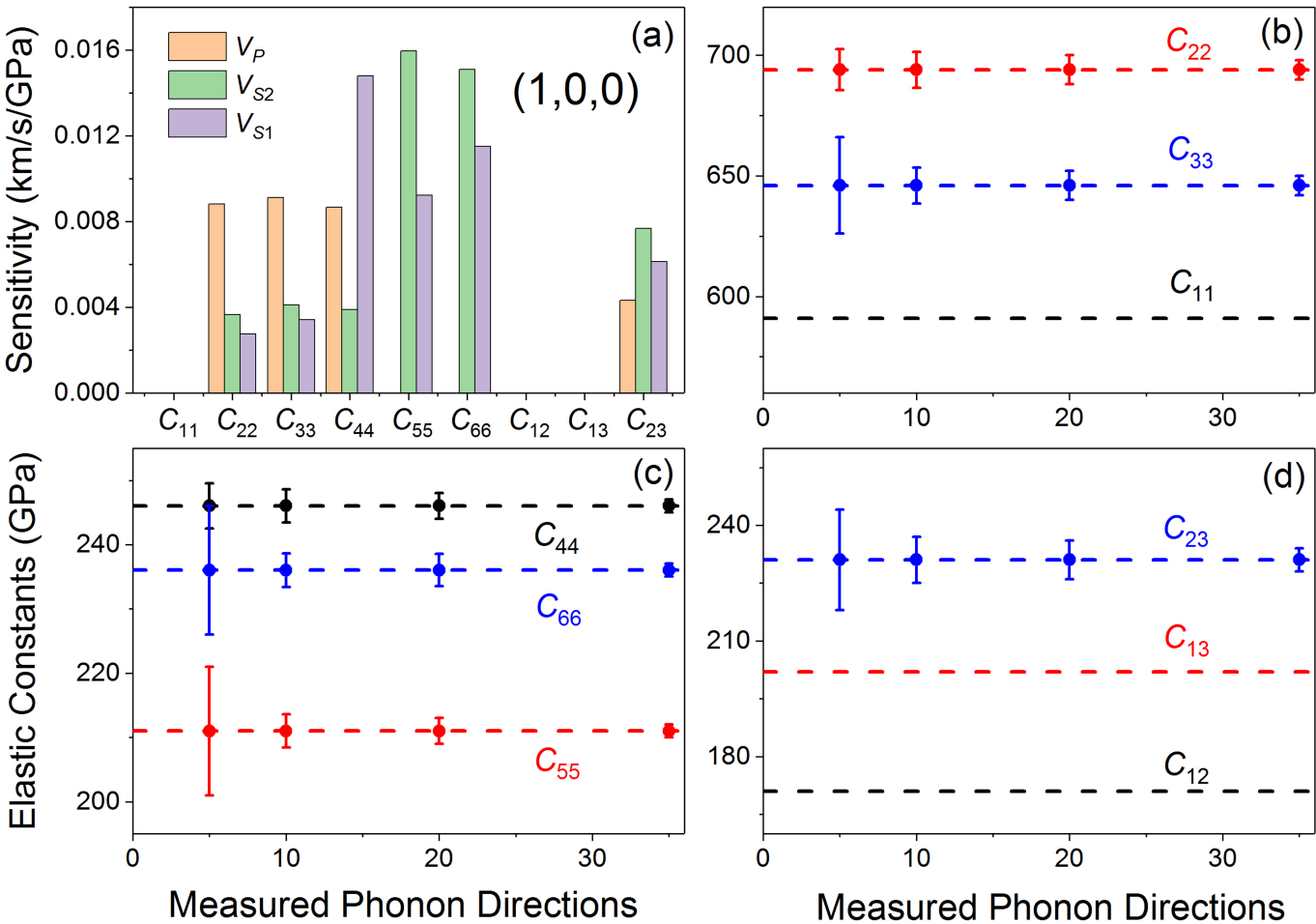
Extended data accompanies this Comment.

Supplementary information accompanies this Comment.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

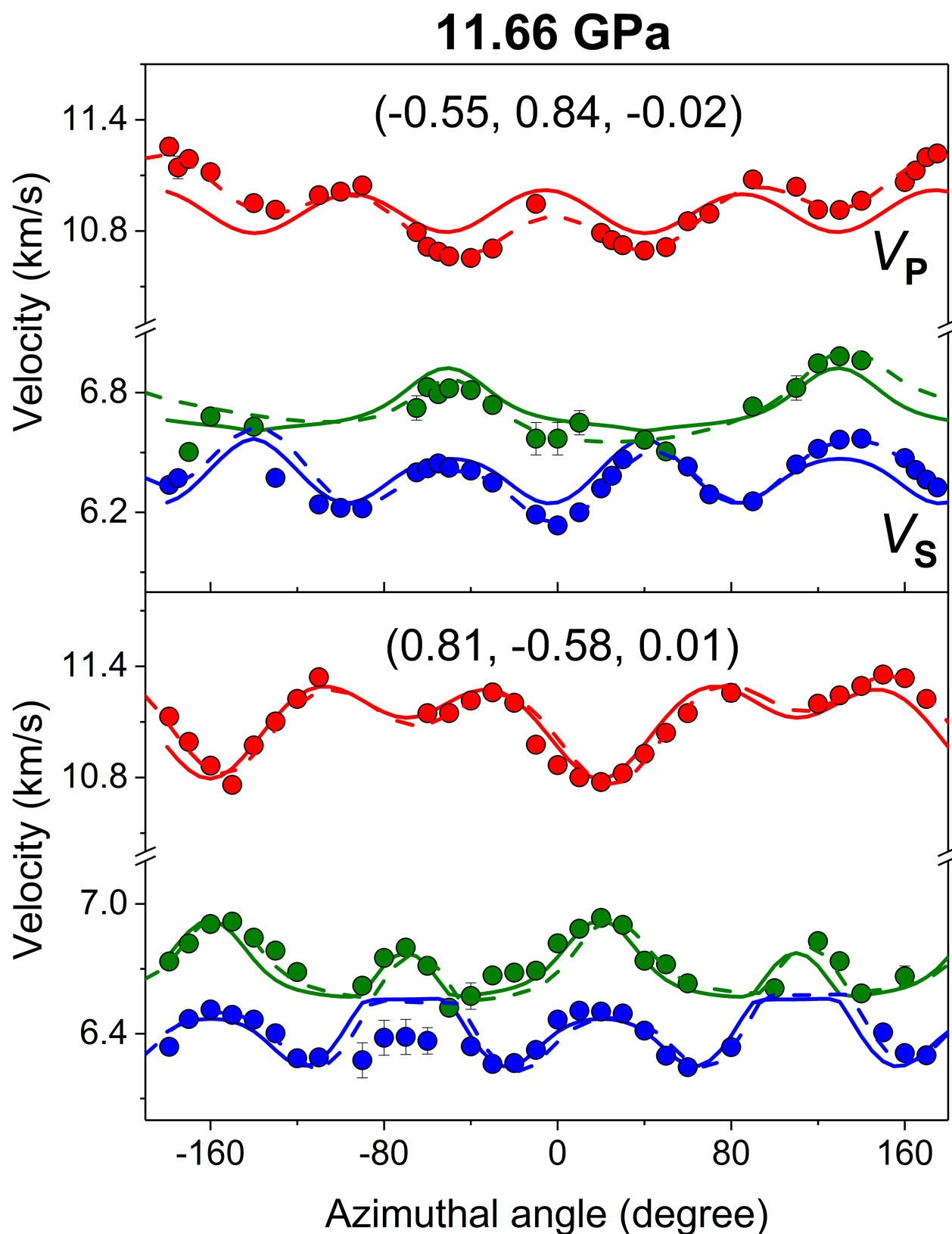
Correspondence and requests for materials should be addressed to J.-F.L.

<https://doi.org/10.1038/s41586-018-0741-7>



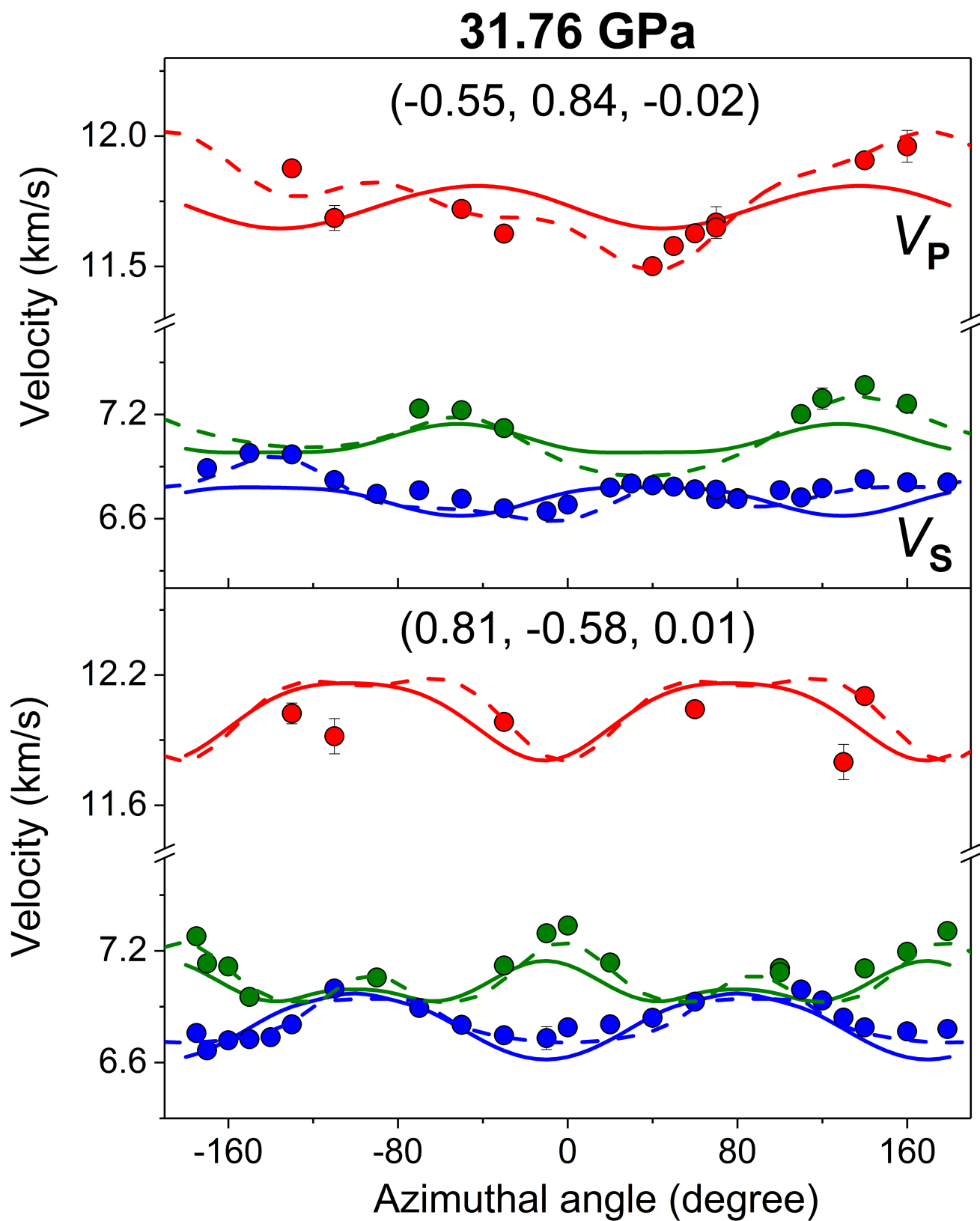
Extended Data Fig. 1 | Relationship between the uncertainties of the reported elastic constants and our calculated sensitivity of wave velocities for a chosen crystallographic (100) orientation as well as the number of measured phonon directions. a–d, The azimuthal

angle between two adjacent phonon directions is 10°. Error bars in b–d represent standard deviations (1σ) of the elastic constants. Data on the elastic constants are from ref. ⁶.



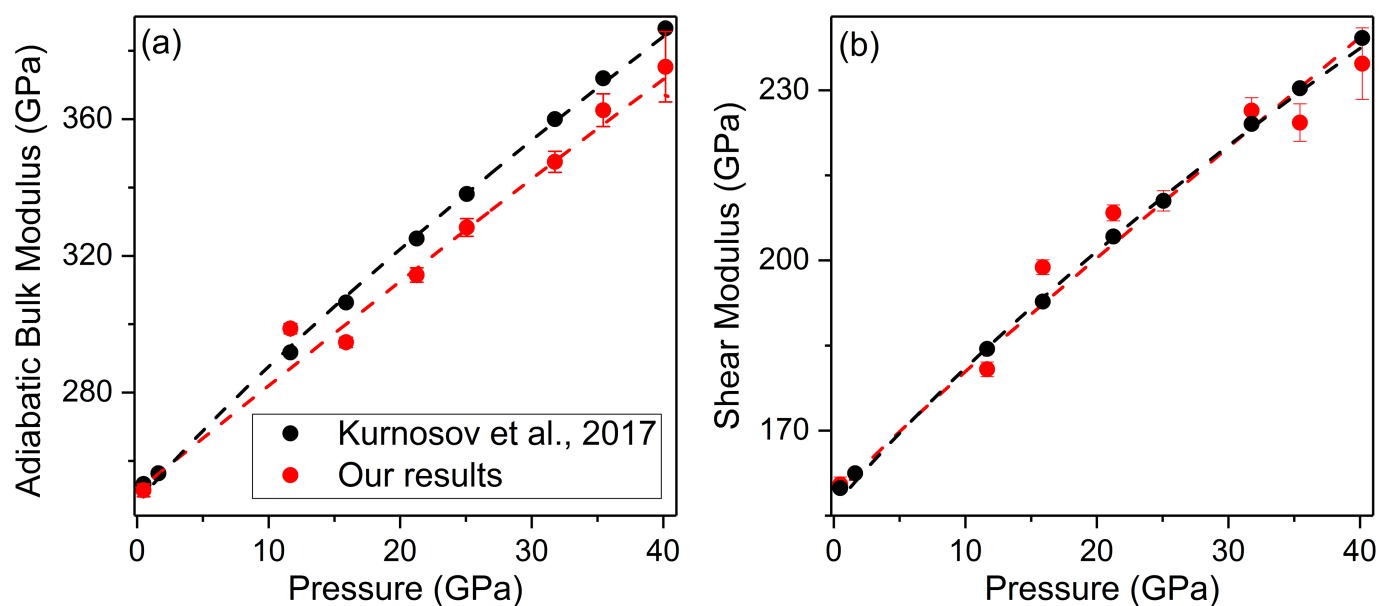
Extended Data Fig. 2 | Modelling the acoustic velocity of bridgmanite to derive C_{ij} at 11.66 GPa. Error bars are not shown when smaller than the symbols. Filled data points are the velocity data from Kurnosov et al.³; the dashed lines are velocity-fitting curves from Kurnosov et al.³; and

the solid lines are velocity-fitting curves using the C_{ij} derived from the covariance matrix analysis in this study. Red, V_P ; green, V_{S1} ; blue, V_{S2} . The orientations of the crystal platelets are given in parentheses.



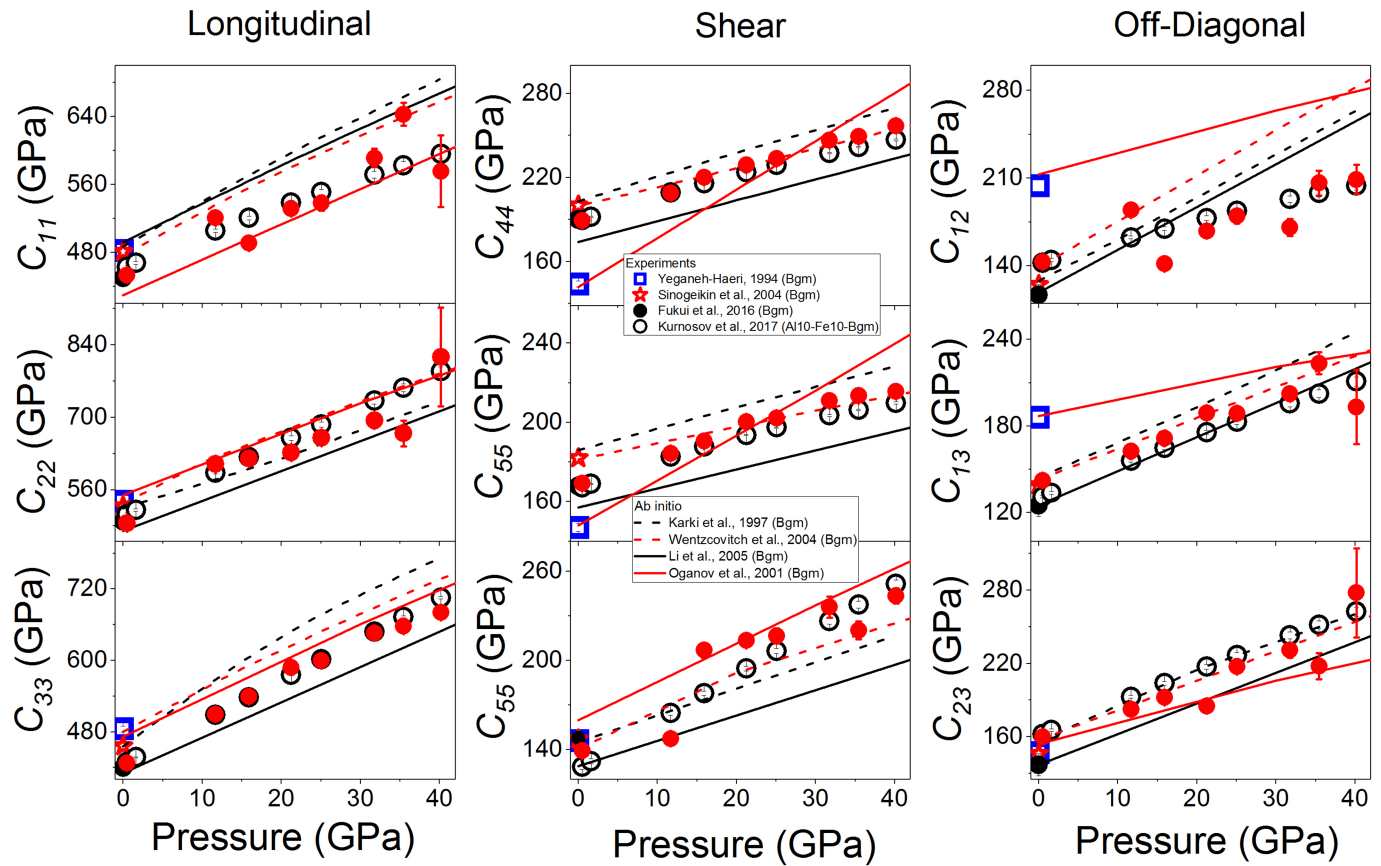
Extended Data Fig. 3 | Modelling the acoustic velocity of bridgmanite to derive C_{ij} at 31.76 GPa. Vertical lines represent standard deviations ($+1\sigma$) and are not shown when smaller than symbols. Solid data points are velocity data from Kurnosov et al.³; the dashed lines are velocity-fitting

curves from Kurnosov et al.³; and the solid lines are velocity-fitting curves using the C_{ij} derived from the covariance matrix analysis in this study. Red, V_P ; green, V_{S1} ; blue, V_{S2} . The orientations of the crystal platelets are given in parentheses.



Extended Data Fig. 4 | Adiabatic bulk modulus and shear modulus of bridgmanite ($\text{Mg}_{0.9}\text{Fe}_{0.1}\text{Si}_{0.9}\text{Al}_{0.1}\text{O}_3$) at high pressure. a, K_S ; b, G . Red data points are results using our obtained full elastic constants (C_{ij}); black data points are data reported in Kurnosov et al.³. Black and red dashed lines are the best fits to data from Kurnosov et al.³ and this study, respectively. Error bars represent standard deviations (1σ) and are not

shown when smaller than the symbols. In this study, the adiabatic bulk modulus at ambient conditions (K_{S0}) is 250(1) GPa, with the pressure derivative of K_S at 300 K (K'_S) = 3.2(2), while the shear modulus at ambient conditions (G_0) is 159(1) GPa, with the pressure derivative of G_0 at 300 K (G') = 2.2(1).

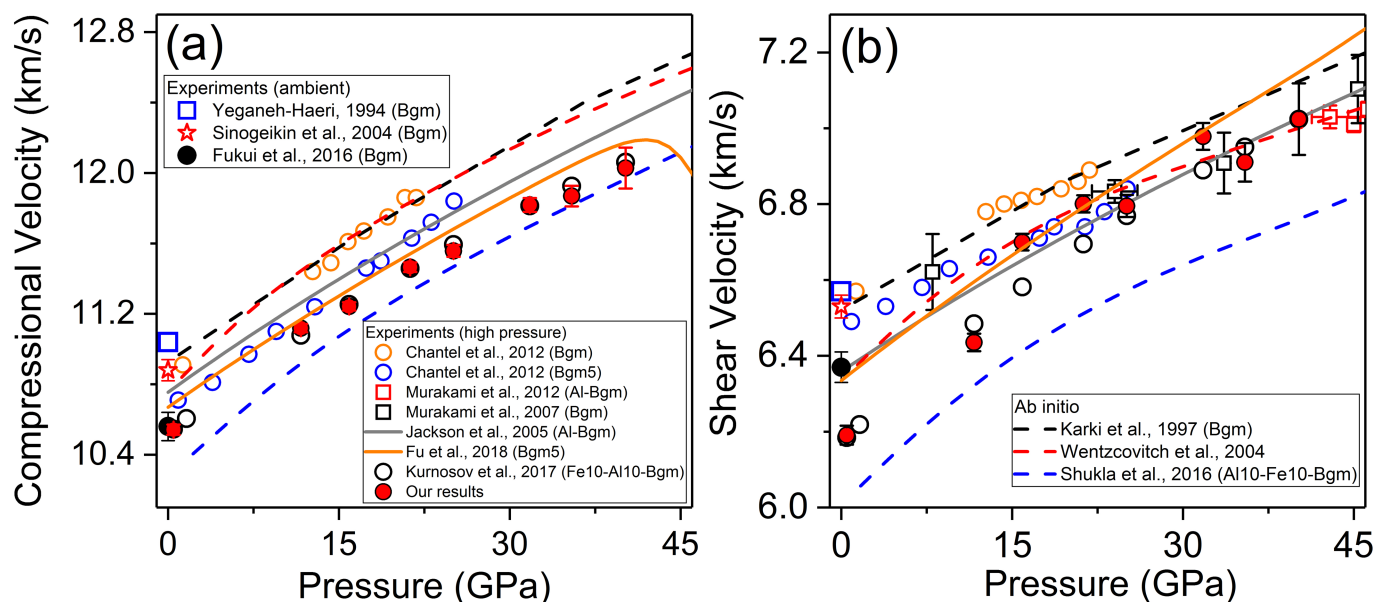


Extended Data Fig. 5 | Comparison of elastic constants of single-crystal bridgmanite as a function of pressure. Bgm, MgSiO_3 bridgmanite; Fe10-Al10-Bgm, (Al,Fe)-bearing bridgmanite with a composition of $(\text{Mg}_{0.9}\text{Fe}_{0.1}\text{Si}_{0.9}\text{Al}_{0.1})\text{O}_3$. Error bars represent standard deviations (1σ)

and are not shown when smaller than the symbols. Symbols indicate experimental results from the literature; lines indicate theoretical calculations^{3,6,7,16–20}. The filled red circles represent the derived elastic constants in this study using the raw velocity data in Kurnosov et al.³.

16. Yeganeh-Haeri, A. Synthesis and re-investigation of the elastic properties of single-crystal magnesium silicate perovskite. *Phys. Earth Planet. Inter.* **87**, 111–121 (1994).
17. Fukui, H. et al. Effect of cation substitution on bridgmanite elasticity: a key to interpret seismic anomalies in the lower mantle. *Sci. Rep.* **6**, 33337 (2016).

18. Karki, B. et al. Elastic properties of orthorhombic MgSiO_3 perovskite at lower mantle pressures. *Am. Mineral.* **82**, 635–638 (1997).
19. Li, L. et al. Elasticity of $(\text{Mg,Fe})(\text{Si,Al})\text{O}_3$ -perovskite at high pressure. *Earth Planet. Sci. Lett.* **240**, 529–536 (2005).
20. Oganov, A. R., Brodholt, J. P. & Price, G. D. Ab initio elasticity and thermal equation of state of MgSiO_3 perovskite. *Earth Planet. Sci. Lett.* **184**, 555–560 (2001).



Extended Data Fig. 6 | Comparison of aggregate compressional and shear wave velocities of single-crystal and polycrystalline bridgmanite at high pressure. Symbols and solid lines represent experimental results from the literature; dashed lines indicate theoretical calculations^{3,6–11,16–18,21,22}. The solid red circles are the calculated velocities

of (Al,Fe)-bearing bridgmanite using the elastic constants derived in this study (Extended Data Fig. 5), while open black circles are from Kurnosov et al.³. Error bars represent standard deviations (1σ) and are not shown when smaller than the symbols.

21. Chantel, J., Frost, D. J., McCammon, C. A., Jing, Z. C. & Wang, Y. B. Acoustic velocities of pure and iron-bearing magnesium silicate perovskite measured to 25 GPa and 1200 K. *Geophys. Res. Lett.* **39**, L19307 (2012).

22. Shukla, G., Cococcioni, M. & Wentzcovitch, R. M. Thermoelasticity of Fe^{3+} - and Al-bearing bridgmanite: effects of iron spin crossover. *Geophys. Res. Lett.* **43**, 5661–5670 (2016).

Kurnosov et al. reply

REPLYING TO J.-F. Lin, Z. Mao, J. Yang & S. Fu *Nature* **564**, <https://doi.org/10.1038/s41586-018-0741-7> (2018)

In our Letter¹, we reported elastic properties for single crystals of Earth's most abundant mineral, (Al,Fe)-bearing bridgmanite, at pressures of the lower mantle, measured by simultaneous Brillouin spectroscopy and X-ray diffraction. We used our data together with previously published results to model seismic wave velocities for Earth's lower mantle. Contradicting previous work², we showed that seismic wave velocities derived for a pyrolitic mantle containing (Al,Fe)-bearing bridgmanite are consistent with the seismic record in the lower mantle up to a depth of about 1,200 km. In the accompanying Comment, Lin et al.³ claim that our measurements are problematic because we measured insufficient velocity data and used insensitive crystal orientations to constrain the full elastic tensor and hence conclude that the uncertainties given in ref. ¹ are underestimated. Here, we address their concerns³ and show them to be unwarranted.

In our Letter¹, we used a novel experimental approach to reduce uncertainties in the derived elastic constants, which consists of: (a) measurement of two focused-ion-beam-prepared crystals^{4,5} of bridgmanite with different crystallographic orientations loaded in a single diamond anvil cell; (b) the simultaneous measurement of acoustic wave velocities and X-ray diffraction determination of the density; and (c) the implementation of a routine to fit the obtained data simultaneously by combining the Christoffel equation with the well established finite strain formalism, which we refer to as a 'global fit'.

In particular, Lin et al.³ independently analysed our experimental data and obtained uncertainties that are different from those derived from our 'global fit'. Lin et al.³ employed a standard approach to extract elastic constants from the raw velocity data that we shared with them; that is, velocities measured at a single pressure have been fitted to the Christoffel equation⁶. The uncertainties that Lin et al.³ report are therefore based on their fit to a single pressure point as opposed to the entire dataset.

In our Letter¹, we measured single-crystal X-ray diffraction of the crystal platelets at each pressure and therefore we know the unit cell volume and the exact crystallographic orientations of both crystals at each pressure. Since the elastic constants at different pressures are not independent but are functions of strain⁷, which was derived in our study from experimentally in situ measured unit cell volumes at every pressure, we implemented a routine to fit all the data simultaneously by combining the Christoffel equation with a finite strain formalism (equation (31) in ref. ⁷), as stated in our Letter¹. In this 'global fit', we simultaneously account for: (a) all measured acoustic wave velocities at all pressures; (b) the X-ray orientations measured for each crystal at each pressure; and (c) the measured unit cell volumes at each pressure.

In other words, rather than fitting all elastic constants independently at each pressure, we fit the parameters of the finite strain equation—which describes the change in the elastic constants with pressure—to all data points at all pressures simultaneously. The 'global fit' greatly decreases the number of fitting parameters if enough pressure points are available. In the standard routine, nine independent elastic constants are refined by the data measured at one pressure point, leading to a ratio between observables and refined parameters of about 10 for each individual pressure. In the 'global fit' only nine zero-pressure elastic constants and pressure derivatives (so 18 parameters in total) need to be refined for the entire dataset and these are constrained by all measurements at all nine pressures. In our Letter¹, approximately 1,000 individual velocities were measured in total and the ratio of observables

to refined parameters therefore increases to about 50. As a result, the 'global fit' substantially reduces the uncertainties in the derived elastic constants, as shown in refs ^{1,8}.

As pointed out by Lin et al.³, the chosen crystallographic orientations have different sensitivities to the different elastic constants. In our Letter¹, the reduced sensitivities of certain constants are taken into account and are reflected in larger uncertainties in the respective values. In all our analyses, cross-correlations between the elastic constants are accounted for by an error propagation and correlation analysis performed as part of the 'Origin 2015 (academic)' fitting routine. The reported error in elastic constants is based on the uncertainties in measured velocities and a propagation of fitting uncertainties and correlations of all fitted parameters. It therefore provides a robust measure of uncertainties that captures the limits imposed by the chosen crystallographic directions, limited velocity coverage at high pressures as well as correlation effects (represented by a covariance matrix, as discussed by Lin et al.³). As stated in our Letter¹, we also analysed our experimental data using the standard approach and find the results to be consistent with the 'global fit', but with substantially larger uncertainties (Fig. 1). Both fits reproduce the measured velocity curves well (Extended Data Fig. 1).

To quantitatively show the effect of employing the 'global fit' on parameter uncertainties, we carried out a series of robustness tests. We used the elastic constants reported in our Letter¹ to generate a synthetic dataset of velocity distribution for the two crystal platelets used in our experiments. Velocities along random crystallographic orientations were then removed to simulate a lack of coverage caused by peak overlapping with the diamond at high pressure or inefficient elasto-optic coupling. Following this procedure, six datasets with limited data coverage were generated that contain between 4 and 36 measured phonon directions; see also Extended Data Fig. 2. These synthetic datasets

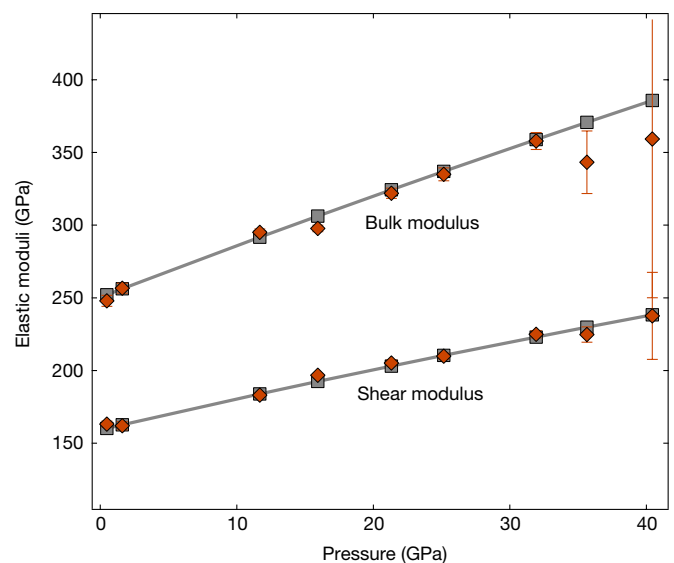


Fig. 1 | Bulk and shear moduli obtained using the 'global fit' (grey squares, solid curves) in comparison to the results of fitting every pressure point individually (red diamonds). Error bars correspond to 2σ . Both methods agree within uncertainties.

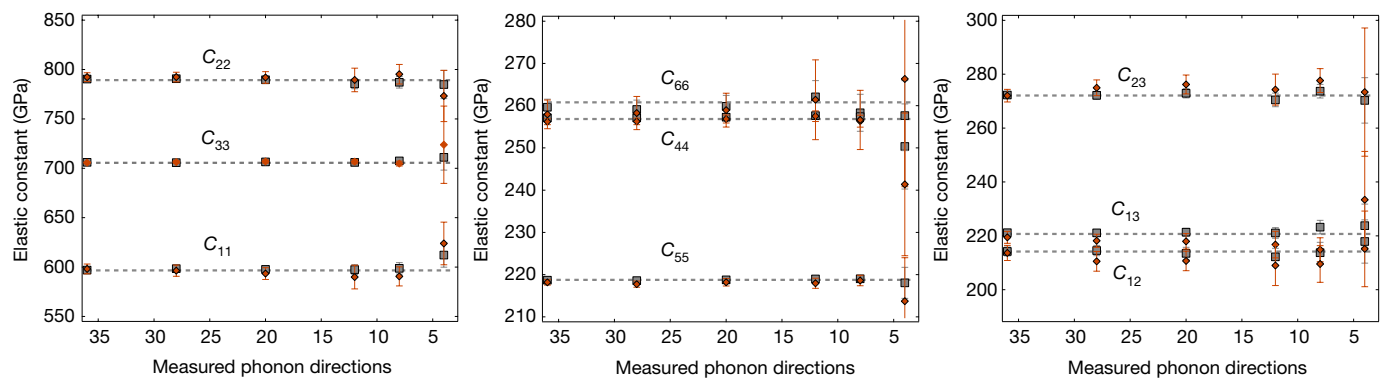


Fig. 2 | Elastic constants derived at 40.4 GPa by fitting the synthetic data. Grey squares are the results of the ‘global fit’; red diamonds are the results of individual fits. Error bars refer to 1σ . The individual fits show systematically larger errors and start to deviate from the real values (dotted lines) when data coverage is limited. The results of the ‘global fit’ stay close to the real values even when few phonon directions are measured. We note

were then analysed in three ways: (a) using the standard approach, where all elastic constants and the orientation of the crystal planes are refined for every pressure individually, as has been done in previous Brillouin studies, for example, refs ^{9–11}; (b) as above, but fixing the orientation of the crystal planes to the values measured by X-ray diffraction, referred to as an ‘individual fit’ here; (c) using the ‘global fit’, where all synthetic velocity data and X-ray volumes are simultaneously inverted to derive the zero-pressure elastic constants and their pressure derivatives.

We found that the standard approach produces the same results and similar uncertainties as the other two routines if data coverage is very good. It fails, however, to refine the elastic constants when fewer velocities are measured (the threshold in our test was about 20 phonon directions). Once the orientations of the crystal planes are known, however, from in situ X-ray diffraction, fewer velocity measurements are required to constrain all elastic constants (Fig. 2, red diamonds). Finally, the application of the ‘global fit’ leads to a significant reduction in uncertainties and reproduces almost all the initial elastic constants at the highest experimental pressure within 1σ , where σ is the standard error on the respective fitting parameter as given by the fitting procedure in the program ‘Origin’ (grey squares in Fig. 2). Figure 3 shows the deviation of all ‘global fit’ elastic constants from the values used to

generate our synthetic dataset (‘real’ values) against their respective ‘global fit’ fitting uncertainties. The deviation of almost all (or all) elastic constants from the real value is smaller than the reported 1σ (or 2σ).

Our tests using the synthetic data show both that the global fit allows the elastic constants to be well constrained even when some gaps in velocity data appear at the highest pressures and that reported fitting uncertainties are accurate. We thus feel that the concerns raised by Lin et al.³ are unwarranted.

Lin et al.³ conclude their Comment by stating that the uncertainties in our data preclude any inference about (a) the existence of metallic iron in the lower mantle or (b) a possible change of ferric iron content with depth in bridgmanite. Although we entirely agree with assertion (a) and have never argued for the possibility of resolving the presence of metallic iron directly from our modelled velocity data (see also figure 3 in ref. ¹), we feel that a discussion about assertion (b) is warranted. There is a subtle, but systematic, deviation in modelled wave velocities from PREM, which is difficult to explain even when uncertainties increase at higher temperatures. In our Letter¹, we discuss the possibility that this deviation reduces if the bridgmanite ferric iron content decreases with depth. Given the importance of such a scenario for our understanding of deep-mantle geophysics and geochemistry, we strongly feel that this finding warrants a discussion and will guide future research. In fact, a recent publication using a complementary approach has independently reported evidence for a decrease in the ferric iron content of the lower mantle with depth¹².

Data availability

The datasets generated and analysed during this study and the Origin routine that we used are available from the corresponding author on reasonable request.

A. Kurnosov¹, H. Marquardt^{1,2*}, D. J. Frost¹, T. Boffa Ballaran¹ & L. Ziberna¹

¹Bayerisches Geoinstitut BGI, University of Bayreuth, Bayreuth, Germany.

²Department of Earth Sciences, University of Oxford, Oxford, UK.

*e-mail: Hauke.Marquardt@uni-bayreuth.de 19 December 2018

1. Kurnosov, A., Marquardt, H., Frost, D. J., Boffa Ballaran, T. & Ziberna, L. Evidence for a Fe³⁺-rich pyrolytic lower mantle from (Al,Fe)-bearing bridgmanite elasticity data. *Nature* **543**, 543–546 (2017); Author Correction *Nature* **558**, E3 (2018).
2. Murakami, M., Ohishi, Y., Hirao, N. & Hirose, K. A perovskitic lower mantle inferred from high-pressure, high-temperature sound velocity data. *Nature* **485**, 90–94 (2012).
3. Lin, J.-F., Mao, Z., Yang, J. & Fu, S. Elasticity of lower-mantle bridgmanite. *Nature* **564**, <https://doi.org/10.1038/s41586-018-0741-7> (2018).

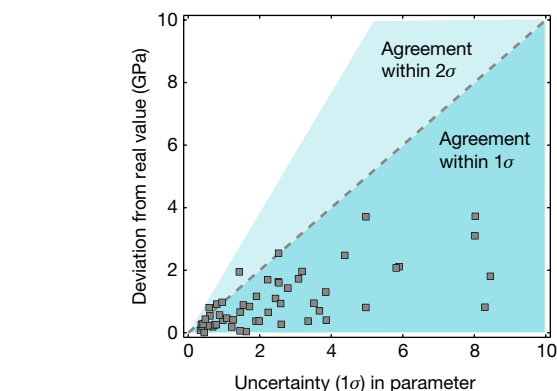


Fig. 3 | Uncertainty analysis for the ‘global fit’. Deviation of all ‘global fit’ elastic constants at 40.4 GPa (that is, all constants derived from all six models) from the values used to generate the dataset (the ‘real’ values) as a function of the fitting uncertainties calculated by the least-squares routine. Almost all derived constants agree with the real values within 1σ (shown as the dark-blue region, that is, all data below the dotted grey line); all constants agree within 2σ (shown as the light-blue region).

4. Marquardt, H. & Marquardt, K. Focused ion beam preparation and characterization of single-crystal samples for high-pressure experiments in the diamond-anvil cell. *Am. Mineral.* **97**, 299–304 (2012).
5. Schulze, K., Buchen, J., Marquardt, K. & Marquardt, H. Multi-sample loading technique for comparative physical property measurements in the diamond-anvil cell. *High Press. Res.* **37**, 159–169 (2017).
6. Speziale, S., Marquardt, H. & Duffy, T. S. Brillouin scattering and its application in geosciences. *Rev. Miner. Geochem.* **78**, 543–603 (2014).
7. Stixrude, L. & Lithgow-Bertelloni, C. Thermodynamics of mantle minerals—I. Physical properties. *Geophys. J. Int.* **162**, 610–632 (2005).
8. Buchen, J. et al. High-pressure single-crystal elasticity of wadsleyite and the seismic signature of water in the shallow transition zone. *Earth Planet. Sci. Lett.* **498**, 77–87 (2018).
9. Speziale, S. & Duffy, T. S. Single-crystal elasticity of fayalite to 12 GPa. *J. Geophys. Res.* **109**, B12202 (2004).
10. Speziale, S. & Duffy, T. S. Single-crystal elastic constants of fluorite (CaF₂) to 9.3 GPa. *Phys. Chem. Miner.* **29**, 465–472 (2002).
11. Marquardt, H., Speziale, S., Reichmann, H. J., Frost, D. J. & Schilling, F. R. Single-crystal elasticity of (Mg_{0.9}Fe_{0.1})O to 81 GPa. *Earth Planet. Sci. Lett.* **287**, 345–352 (2009).
12. Shim, S.-H. et al. Stability of ferrous-iron-rich bridgmanite under reducing midmantle conditions. *Proc. Natl Acad. Sci. USA* **114**, 6468–6473 (2017).
13. Sinogeikin, S. V. & Bass, J. D. Single-crystal elasticity of pyrope and MgO to 20 GPa by Brillouin scattering in the diamond cell. *Phys. Earth Planet. Inter.* **120**, 43–62 (2000).

Author contributions A.K., H.M., D.J.F. and T.B.B. discussed the content. A.K. performed the robustness test. H.M. wrote the paper draft. All authors commented on the manuscript.

Competing interests Declared none.

Additional information

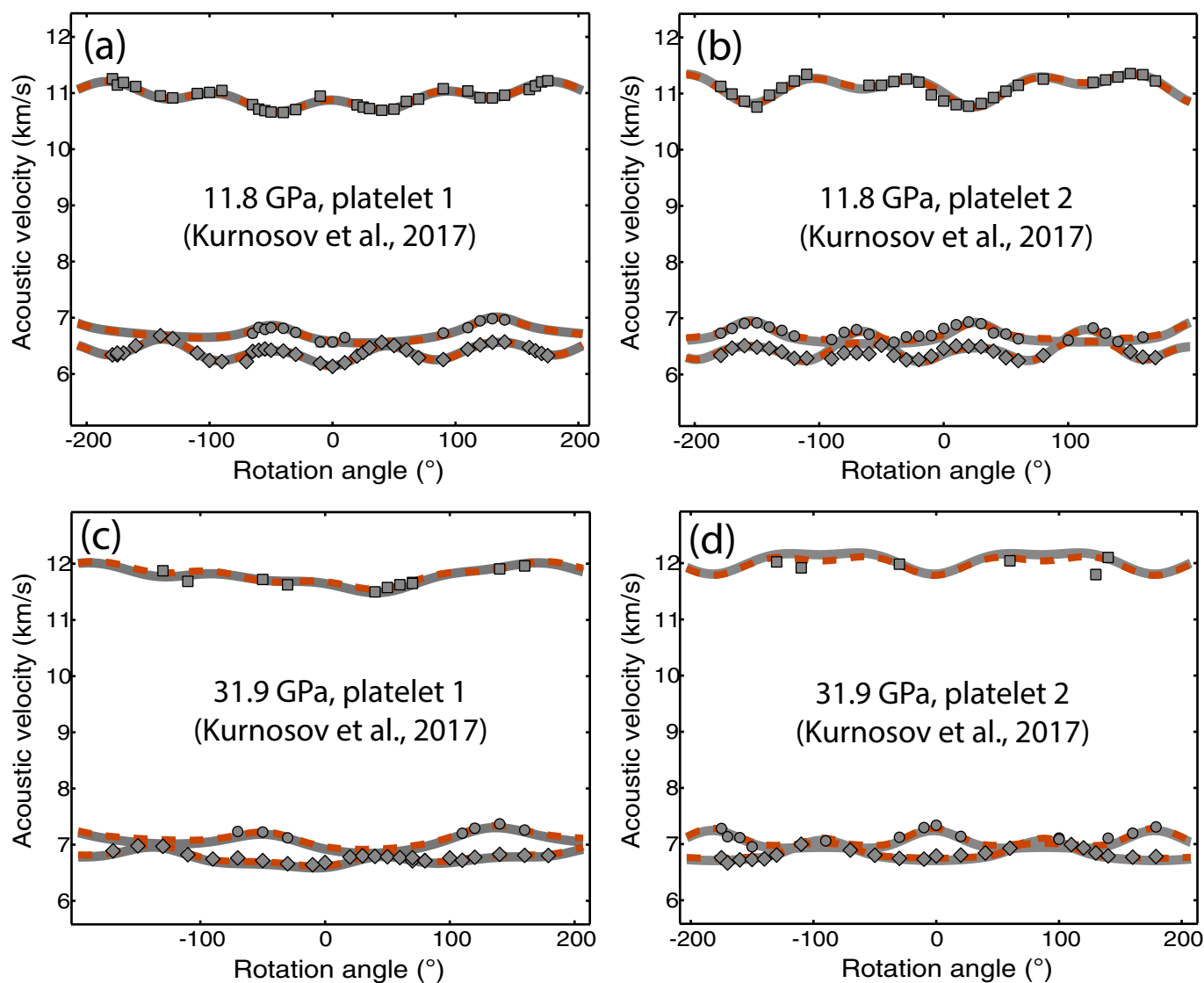
Extended data accompanies this Reply.

Supplementary information accompanies this Reply.

Reprints and permissions information is available at <https://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to H.M.

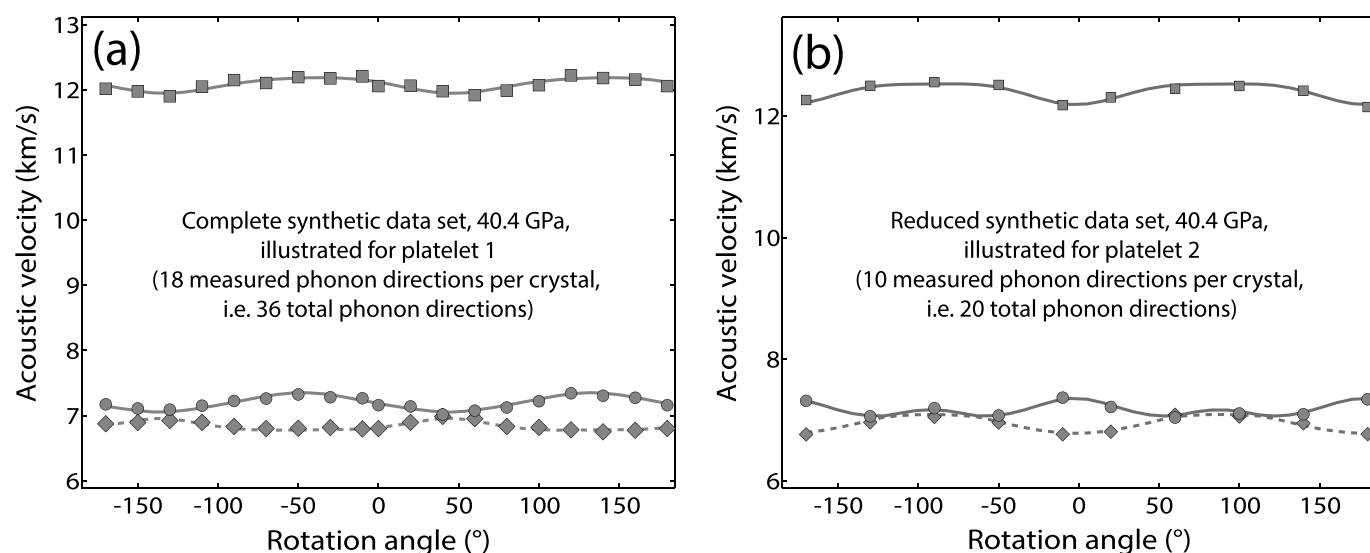
<https://doi.org/10.1038/s41586-018-0742-6>



Extended Data Fig. 1 | Raw data collected on single-crystal bridgmanite at two selected pressures and fits from our Letter. The solid curves are fits to the grey data using the best-fit global model (the ‘global fit’ of ref. ¹) and the red dashed curves are individual fits. **a, c**, Platelet 1. **b, d**, Platelet 2. **a, b**, Data collected at 11.8 GPa; **c, d**, data collected at 31.9 GPa. We note

that the orientation of the crystals changed slightly during pressure increase, but was measured at every pressure in our experiment. A ‘tilt’ correction was employed for crystal platelet 1, as mentioned in our Letter¹. The procedure is further explained in the Supplementary Information (containing the raw data).

BRIEF COMMUNICATIONS ARISING



Extended Data Fig. 2 | Illustration of how the synthetic dataset at 40.4 GPa was generated using the elastic constants reported in our Letter. For every pressure point and platelet, velocities in 18 propagation directions were calculated over an angular range of 360° to simulate in-plane rotations between individual measurements that are representative of our low pressure measurements where no overlap of peaks occurred (solid symbols). We randomly changed the velocity values by up to 0.5% to simulate errors typical for single velocity measurements by Brillouin

spectroscopy in the diamond anvil cell¹³. **a**, Illustration of a complete velocity dataset, using crystal platelet 1 as an example. **b**, Illustration of a reduced dataset, using crystal platelet 2 as an example. This assumes limited data availability owing to peak overlap or elasto-optic coupling. The example in **b** corresponds to a situation where 20 directions have been measured in total (10 on each crystal). Solid and dotted curves are results from the 'global fit'.

Concerns about modelling of the EDGES data

ARISING FROM J. D. Bowman, A. E. E. Rogers, R. A. Monsalve, T. J. Mozdzen & N. Mahesh *Nature* **555**, 67–70 (2018); <https://doi.org/10.1038/nature25792>

It is predicted¹ that the spectrum of radio emission from the whole sky should show a dip arising from the action of the light from the first stars on the hydrogen atoms in the surrounding gas, which causes the 21-cm line to appear in absorption against the cosmic microwave background. Bowman et al.² identified a broad flat-bottomed absorption profile centred at 78 MHz, which could be this feature, although the depth of the profile is much larger than expected. We have examined the modelling process they used and find that their data implies unphysical parameters for the foreground emission and also that their solution is not unique, in the sense that we found other simple formulations for the signal that are different in shape but that also fit their data. We argue that this calls into question the interpretation

of these data as an unambiguous detection of the cosmological 21-cm absorption signature. There is a Reply to this Comment by Bowman, J. D. et al. *Nature* **564**, <https://doi.org/10.1038/s41586-018-0797-4> (2018).

Bowman et al.² describe a ‘physically motivated’ foreground model containing three parameters describing synchrotron emission (magnitude, spectral index and the ‘running’ of the index) and two for ionospheric emission and absorption. They used a linearized version of this model to perform fits with and without the 21-cm feature. Using this model and the data that they released, we obtained essentially identical results, but we note that accommodating the proposed absorption profile requires a change in the foreground model that

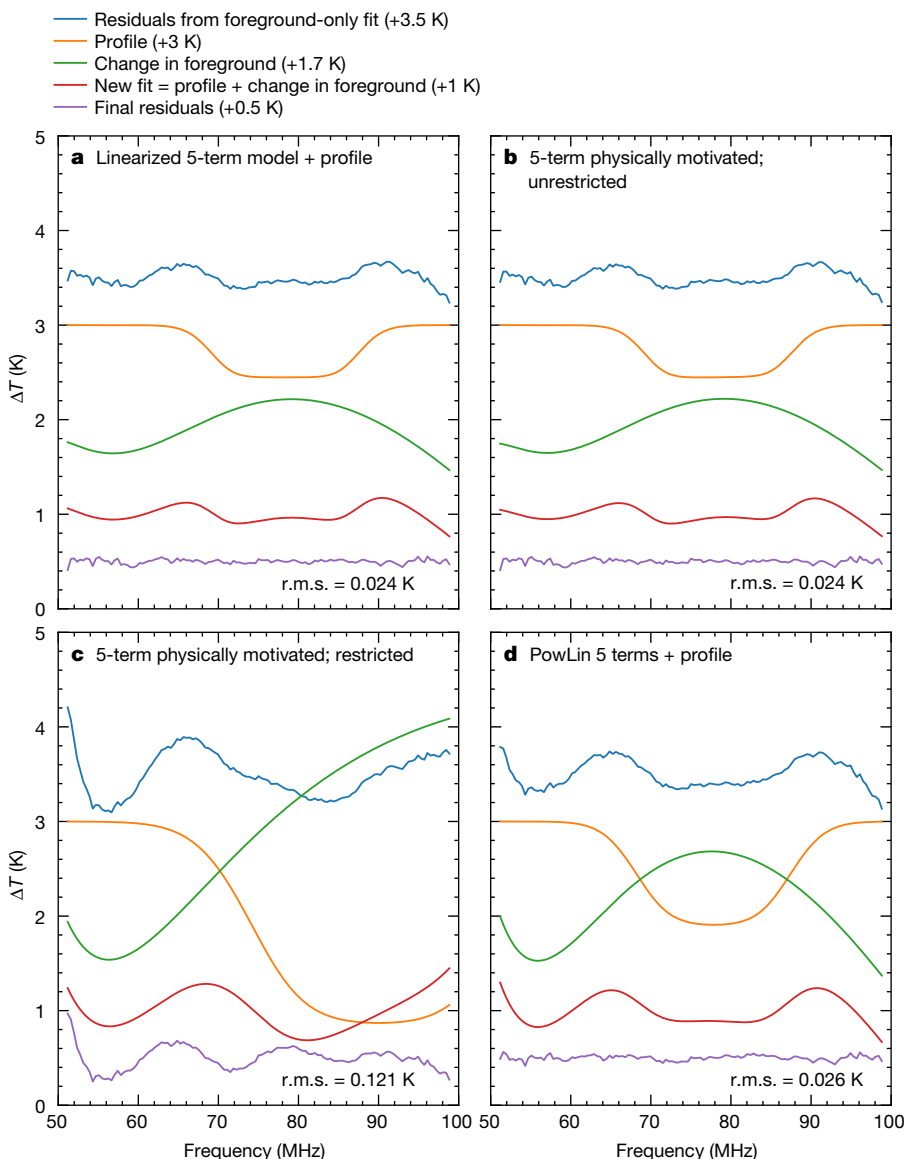


Fig. 1 | Fits to the EDGES data. **a**, With the foregrounds described by the linearized function used by Bowman et al.². **b**, Using the physically motivated nonlinear function for the foregrounds and no restrictions on the parameters. **c**, The same as **b** but with the range of parameters limited to physically plausible values. **d**, Using the PowLin model, which consists of a power law with index given by a polynomial in frequency, ν . The top line in each panel shows the residuals when a fit is made using the foreground model only. The bottom line is the residual when the fit is run again including the profile with the functional form given by Bowman et al.². The intermediate lines show the shape of the profile found, the change in the foreground model needed to accommodate this and the sum of these two, which is also equal to the difference between the initial and final residuals. The curves have been offset vertically for readability. (r.m.s., root mean square.)

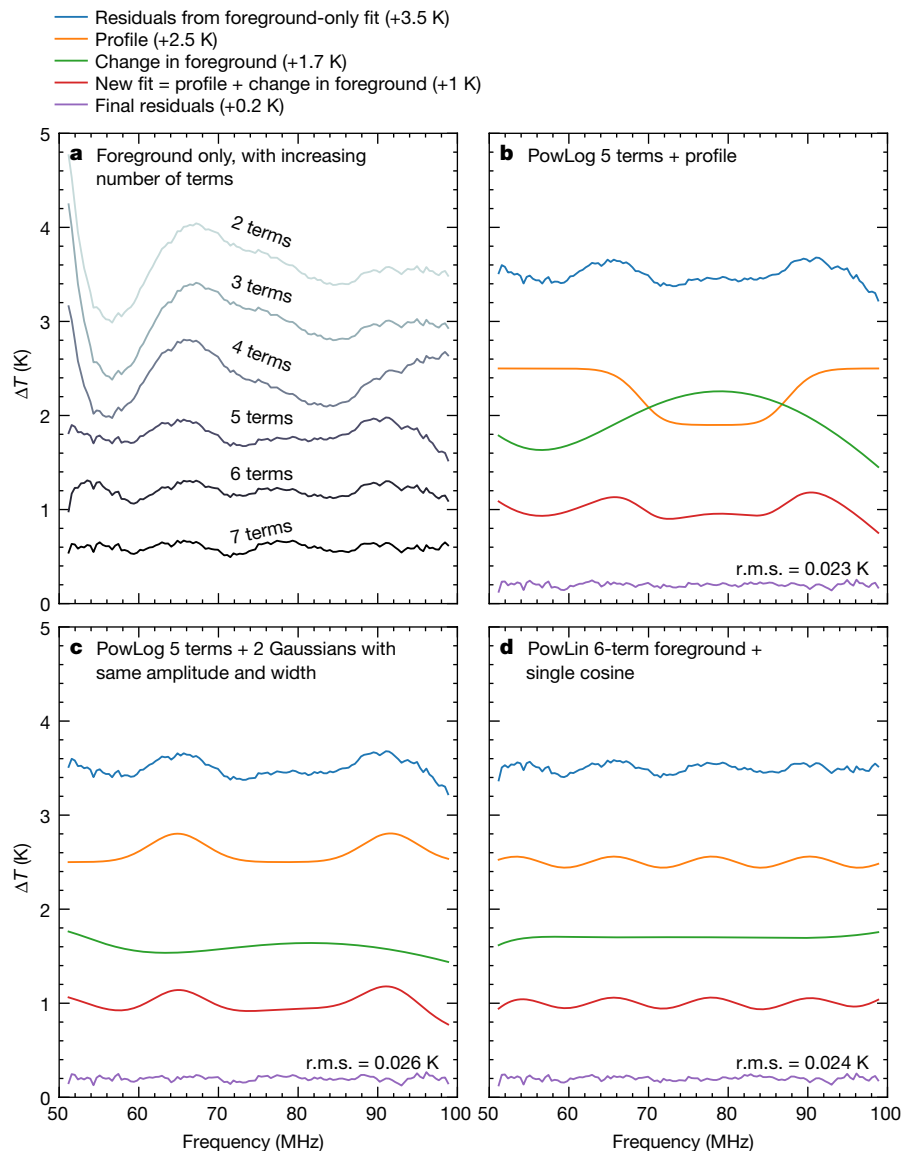


Fig. 2 | Further illustrations of the fitting process. **a**, The residuals when the ionospheric parameters are set to $\tau_0 = 0.014$ and $T_e = 800$ K and the PowLog model—a power law with index given by a polynomial in $\log(\nu)$ —is fitted with increasing numbers of terms. **b**, Showing how the Bowman et al.² profile with four signal parameters can provide a good fit by making use of the freedom provided by five foreground parameters.

c, d, Alternative nine-parameter fits. **c**, The same model as **b** but with two Gaussian features of equal width and amplitude in place of the Bowman et al.² profile. **d**, The PowLin model with six terms and, instead of the profile, a single cosine function, which has a fitted amplitude of about 0.06 K. The different curves have again been offset vertically for readability.

is much larger than the initial residuals; see Fig. 1a. We also found that the parameters describing the foregrounds have unphysical values; for example, the parameter associated with brightness temperature of the ionospheric emission exceeds 10^4 K, while that for the astronomical foreground brightness has a large negative value. Full details of the functions fitted, together with the values found for the parameters, are given in the Supplementary Information to this Comment.

To gain further insight we fitted the full non-linear expression, taking into account the linkage of the emission and absorption by the ionosphere via the temperature of the electrons, T_e . The values found for the optical depth of the ionosphere τ_{ion} and for T_e are both negative, which is clearly unphysical. We constrained these parameters to physically plausible values³, $\tau_{\text{ion}} > 0.005$ at 75 MHz and $200 < T_e < 2,000$ K, and we restricted the centre frequency of the absorption profile to lie

between 60 and 90 MHz. The results obtained with and without these restrictions on the parameters are shown in Fig. 1b, c. Without the restrictions we obtained essentially the same profile as Bowman et al.² and the fit is good. (We describe the fit as ‘good’ whenever the root mean square of the residuals is below 0.03 K.) With the restrictions the fit is poor; the centre of the profile has moved to the upper limit and its depth has increased to about 2 K.

We then explored cases where the ionospheric opacity and temperature are held fixed at reasonable values but higher-order terms are added to the foreground model, using several different formulations. We found that at least five free foreground parameters, in addition to the four absorption profile parameters, were always needed to obtain a good fit. The parameters found for the profile changed substantially when different formulations for the foreground were used. Figure 1d shows an example of this, where a good fit was obtained with an

amplitude of about 1.1 K for the absorption feature, which is even larger than that found by Bowman et al.²

The residuals found when fitting with successively higher numbers of terms in the foreground model are shown in Fig. 2a. We note that at no stage in this process does a distinct absorption-line feature appear. Instead, a broad oscillatory feature is present when two, three or four terms are used and it is only the addition of the fifth term that produces a large reduction in the residuals. Although it can be argued that higher-order terms are needed to represent the synchrotron foreground accurately, this is not the behaviour expected⁴. In particular, the relatively large value required for the fifth term is not consistent with what is known^{5–8} about the spectrum of the foreground emission in the range 25–400 MHz. Adding higher-order foreground terms has simply moved the problem of unphysical parameter values from the ionosphere to the foreground.

It seems possible that the unphysical values are due to residual systematic errors in the data, perhaps arising in the correction for the frequency-dependent beam shape, but if that is the case then it is not clear that model formulations chosen to suit the astronomical foregrounds are the correct way of removing such effects and there is also no clear basis for deciding how many terms should be included. A general concern is that nine parameters are being fitted to data that span 50 MHz and contain very little real structure with periods shorter than about 10 MHz, so neighbouring data points in the spectrum are strongly correlated. This means that the number of truly independent data points may not be much larger than the number of parameters being fitted.

We next demonstrated that the profile found by Bowman et al.² is not a unique solution. The top lines of Fig. 1a, b show the residuals after subtracting a five-parameter foreground fit. As already noted, the ionosphere parameters have unphysical values in those cases. The fourth line down in Fig. 2a shows that very similar residuals are obtained by assuming reasonable fixed values for the ionosphere and fitting a five-parameter power law. The residuals do not, however, show an absorption profile but instead show two peaks at around 65 MHz and 90 MHz. Although one can obtain a good fit using the Bowman et al.² profile (Fig. 2b), a good fit can also be achieved with two Gaussian emission features of equal height and width (Fig. 2c). With more terms in the foreground model (that is, the bottom lines in Fig. 2a), the residuals take the form of undulations with a period of around 12.5 MHz. We found that a satisfactory fit can then be obtained with just a sine wave, as shown in Fig. 2d. In both of these models the total number of free parameters is again nine.

We also found that, with the 12.5 MHz sine wave removed, a good fit was obtained with five foreground parameters and a broad Gaussian absorption profile and that there is then a large covariance between the foreground and signal components. This suggests that these undulations may be what causes the fitting process used by Bowman et al.² to produce a profile with a flattened bottom. Since the proposed 21-cm absorption profile does not match theoretical expectations in either shape or amplitude, it is not clear why it should be preferred to the other forms of signal explored here or to the many more that can be found in the degenerate space between signal and foreground model. Therefore, although our analysis does not prove that the feature identified by Bowman et al.² is absent from their data, we believe the issues that we have raised are such that the evidence for its presence falls well short of the level required to invoke new physics for its explanation.

G.K. acknowledges support from ERC Advanced Grant 320596 ‘The Emergence of Structure During the Epoch of Reionization’. P.D.M. and E.P. acknowledge support from Senior Kavli Institute Fellowships at the University of Cambridge. P.D.M. also acknowledges support from The Netherlands Organization For Scientific Research (NWO) VIDI grant (dossier 639.042.730).

Methods

We used a least-squares fitting for testing the models presented in the text and the Supplementary Information. In addition, we derived posterior distributions of model parameters by implementing a likelihood function into the multi-nested sampler Polychord^{9,10}.

Data availability

For all our analyses we used the ‘Data for Figure 1 of Bowman et al. (2018)’ in the EDGES Data Release, which is available at <http://loco.lab.asu.edu/edges/edges-data-release/>.

Richard Hills^{1*}, Girish Kulkarni^{2,3,7}, P. Daniel Meerburg^{2,3,4,5,6} & Ewald Puchwein^{2,3,8}

¹Cavendish Laboratory, University of Cambridge, Cambridge, UK.

²Institute of Astronomy, University of Cambridge, Cambridge, UK.

³Kavli Institute for Cosmology, University of Cambridge, Cambridge, UK.

⁴Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK.

⁵Kapteyn Astronomical Institute, University of Groningen, Groningen, The Netherlands.

⁶Van Swinderen Institute for Particle Physics and Gravity, University of Groningen, Groningen, The Netherlands.

⁷Present address: Department of Theoretical Physics, Tata Institute of Fundamental Research, Mumbai, India.

⁸Present address: Leibniz Institute for Astrophysics, Potsdam, Germany.

*e-mail: richard@mrao.cam.ac.uk

Received: 3 May 2018; Accepted: 15 October 2018;

Published online 19 December 2018.

1. Pritchard, J. R. & Loeb, A. 21 cm cosmology in the 21st century. *Rep. Prog. Phys.* **75**, 086901 (2012).
2. Bowman, J. D., Rogers, A. E. E., Monsalve, R. A., Mozdzen, T. J. & Mahesh, N. An absorption profile centred at 78 megahertz in the sky-averaged spectrum. *Nature* **555**, 67–70 (2018).
3. Rogers, A. E. E., Bowman, J. D., Vierinen, J., Monsalve, R. & Mozdzen, T. Radiometric measurements of electron temperature and opacity of ionospheric perturbations. *Radio Sci.* **50**, 130–137 (2015).
4. Bernardi, G., McQuinn, M. & Greenhill, L. J. Foreground model and antenna calibration errors in the measurement of the sky-averaged λ 21 cm signal at $z \sim 20$. *Astrophys. J.* **799**, 90 (2015).
5. Bridle, A. H. The spectrum of the radio background between 13 and 404 MHz. *Mon. Not. R. Astron. Soc.* **136**, 219–240 (1967).
6. de Oliveira-Costa, A. et al. A model of diffuse Galactic radio emission from 10 MHz to 100 GHz. *Mon. Not. R. Astron. Soc.* **388**, 247–260 (2008).
7. Mozdzen, T. J., Bowman, J. D., Monsalve, R. A. & Rogers, A. E. E. Improved measurement of the spectral index of the diffuse radio background between 90 and 190 MHz. *Mon. Not. R. Astron. Soc.* **464**, 4995–5002 (2017).
8. Zheng, H. et al. An improved model of diffuse galactic radio emission from 10 MHz to 5 THz. *Mon. Not. R. Astron. Soc.* **464**, 3486–3497 (2017).
9. Handley, W. J., Hobson, M. P. & Lasenby, A. N. POLYCHORD: nested sampling for cosmology. *Mon. Not. R. Astron. Soc.* **450**, L61–L65 (2015).
10. Handley, W. J., Hobson, M. P. & Lasenby, A. N. POLYCHORD: nested sampling for cosmology. *Mon. Not. R. Astron. Soc.* **453**, 4385–4399 (2015).

Author contributions E.P. initiated this study by pointing out the importance of the role of the foreground parameters in the claimed detection of an absorption profile. All authors participated in the detailed analysis and the writing of the Comment.

Competing interests Declared none.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0796-5>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to R.H.

<https://doi.org/10.1038/s41586-018-0796-5>

Reply to Hills et al.

REPLYING TO R. Hills et al. *Nature* **564**, <https://doi.org/10.1038/s41586-018-0796-5> (2018)

In our Letter¹, the foreground models account for a combination of astronomical foregrounds, ionospheric effects and any residual calibration effects. We obtained best-fit parameters that absorb a degenerate combination of these effects from unconstrained fits to the models. In the accompanying Comment², the concerns of Hills et al. arise primarily because they failed to recover physical values for two ionosphere parameters in a foreground model with three additional non-ionosphere parameters.

Ionosphere parameters are covariant with the amplitude and spectral index of the astronomical foreground. Small errors in these astronomical parameters, as well as residual effects from calibration, could bias the recovered ionosphere parameters. In ref. ³, we calculated an overall systematic uncertainty of ± 0.02 on the spectral index measured by the high-band instrument, including beam correction uncertainty. We found a similar uncertainty for the low-band instrument. Errors of this level could yield deviations from the true spectrum with amplitudes and shapes comparable to those of the expected ionospheric contribution. For these reasons, we did not intend to extract ionospheric information from the measurements presented. In a previous study⁴, we extracted information about ionospheric variability from EDGES high-band data. We limited that analysis to differencing spectra acquired at the same local sidereal time on different nights in order to reduce the covariance with the astronomical foregrounds and mitigate any systematic effects, before fitting an ionosphere model to the differential spectra. Extracting absolute ionospheric information directly from the measured spectra would require a separate, in-depth study.

Measuring physical foreground properties requires the absolute temperature calibration of the spectrum, whereas identifying a 21-cm profile embedded in the foreground requires only relative calibration between channels in the spectrum. It is possible to recover a 21-cm feature without accurately measuring the physical foreground properties. Most global 21-cm constraints have come from this regime^{5–9}. In EDGES we do aim to measure a fully absolutely calibrated spectrum. Although in our Methods section we acknowledged potential residual calibration effects, we reported tests to show that any such effects are not consistent with the reported profile. We therefore concluded that the signal is astronomical.

Hills et al.² found that several alternative models for the foreground and signal can be fitted to the data. We broadly agree, but a general absorption profile remains the most justified a priori choice of signal model because we have disfavoured the instrument as the source of the structure and there is no known physical expectation for other shapes in either the foreground or 21-cm signal, whereas an absorption is expected. We have data that exclude some of the alternative signal models proposed by Hills et al.² and plan to publish those results in the near future.

When using our polynomial foreground model over the full band (51–99 MHz), rather than over only the sub-band for which we used it (approximately 63–99 MHz), Hills et al.² recovered best-fit profiles that are not consistent with our reported properties. We have shown using simulations¹⁰ that this outcome is consistent with the expected

performance of that model. Therefore, their choice to use it over the full band was not justified. Other foreground models perform better than the polynomial model across the full band, including the linear physically motivated model that we used.

Judd D. Bowman^{1*}, Alan E. E. Rogers², Raul A. Monsalve^{1,3,4,5,6}, Thomas J. Mozdzen¹ & Nivedita Mahesh¹

¹School of Earth and Space Exploration, Arizona State University, Tempe, AZ, USA.

²Haystack Observatory, Massachusetts Institute of Technology, Westford, MA, USA.

³Department of Physics, McGill University, Montréal, Quebec, Canada.

⁴McGill Space Institute, McGill University, Montréal, Quebec, Canada.

⁵Center for Astrophysics and Space Astronomy, University of Colorado, Boulder, CO, USA.

⁶Facultad de Ingeniería, Universidad Católica de la Santísima Concepción, Concepción, Chile.

*e-mail: judd.bowman@asu.edu

Published online 19 December 2018.

1. Bowman, J. D., Rogers, A. E. E., Monsalve, R. A., Mozdzen, T. J. & Mahesh, N. An absorption profile centred at 78 megahertz in the sky-averaged spectrum. *Nature* **555**, 67–70 (2018).
2. Hills, R., Kulkarni, G., Meerburg, P. D. & Puchwein, E. Concerns about modelling of the EDGES data. *Nature* **564**, <https://doi.org/10.1038/s41586-018-0797-4> (2018).
3. Mozdzen, T. J., Bowman, J. D., Monsalve, R. A. & Rogers, A. E. E. Improved measurement of the spectral index of the diffuse radio background between 90 and 190 MHz. *Mon. Not. R. Astron. Soc.* **464**, 4995–5002 (2017).
4. Rogers, A. E. E., Bowman, J. D., Vierinen, J., Monsalve, R. & Mozdzen, T. Radiometric measurements of electron temperature and opacity of ionospheric perturbations. *Radio Sci.* **50**, 130–137 (2015).
5. Bowman, J. D. & Rogers, A. E. E. Lower limit of $\Delta z > 0.06$ for the duration of the reionization epoch. *Nature* **468**, 796–798 (2010).
6. Bernardi, G. et al. Bayesian constraints on the global 21-cm signal from the cosmic dawn. *Mon. Not. R. Astron. Soc.* **461**, 2847–2855 (2016).
7. Singh, S. et al. First results on the epoch of reionization from first light with SARAS 2. *Astrophys. J. Lett.* **845**, L12 (2017).
8. Singh, S. et al. SARAS 2 constraints on global 21 cm signals from the epoch of reionization. *Astrophys. J. Lett.* **858**, 54 (2018).
9. Price, D. C. et al. Design and characterization of the Large-aperture Experiment to Detect the Dark Age (LEDA) radiometer systems. *Mon. Not. R. Astron. Soc.* **478**, 4193–4213 (2018).
10. Bowman, J. Foreground model selection for signal parameter estimation. EDGES report 122, http://loco.lab.asu.edu/loco-memos/edges_reports/report122.pdf (2018).

Author contributions J.D.B., R.A.M. and A.E.E.R. contributed equally to this Reply. N.M. and T.J.M. provided input and approved the final response.

Competing interests Declared none.

Additional information

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to J.D.B.

<https://doi.org/10.1038/s41586-018-0797-4>

CAREERS

ENROLMENTS Fewer graduate students head to the United States go.nature.com/2ehrjkd

CONFERENCES Attendance boosts chances of co-authorship go.nature.com/2zqumxn

MENTORING Tips and resources for support go.nature.com/mentor

SOLSTOCK/GETTY



A linear route through academia isn't the only way forward. Keeping your mind — and options — open can lead to unexpected success.

ADVICE

Boost your research career

Five scientists offer tips on how to land a rewarding job and make the most of your science.

MARTIJN BIJCKER

Overhaul your CV for jobs in industry

Immunologist and founder of the training company From Science to Pharma in Sydney, Australia.

Academic scientists who want a job in industry need to get into an industrial mindset, and that starts with the right CV. I've seen it time and time again: academics use the same CV whether they're applying for a job at a pharmaceutical company or an assistant professorship at a university. That's no way to get hired. If you've sent out 50 CVs without a response, you

need to rethink your approach.

One of the biggest mistakes academics make is listing every single thing they've ever done. I've seen CVs that were 17 or 18 pages long, and it's completely unnecessary. If you have a lot of papers, it's good to mention that, but you should list only the most relevant four or five.

Companies aren't especially interested in your papers and poster presentations. They want to know that you can work with a team to get results. To stand out, you should emphasize your collaborations, not your citations. Have you ever served on a committee? Organized a conference or a journal club? Have you done anything that showed initiative and cooperation beyond the lab? Put it at the top of the CV.

Industry is all about outcome, so be sure to highlight your results. You didn't just organize a conference, you organized a highly successful

conference that attracted 20 speakers from around the world. If you don't spell out your success, potential hirers might suspect that you messed everything up.

NAIL THE INTERVIEW

A good CV can easily land you an interview. But if you get a rejection letter that cites your lack of industry experience, it's fair to say that the interview didn't go well. The most likely explanation: you looked fine on paper, but you didn't seem to know what you were talking about in person.

Preparing for an interview takes work: you can't just go in there and wing it, telling them you're a people person and a fast learner. They've heard that before. You have to talk beforehand to people who are doing or have done the job so that you can get into their mindset. The realities of a job can be very different from the ►

► description. The interviewers need to know that you're really invested in the position.

Bijker contributed to the feature 'How to sail smoothly from academia to industry' (*Nature* **555**, 549–551; 2018).

PHILIPP KRUGER Take the initiative

Immunologist and career outreach fellow at the University of Oxford, UK

If you go along doing what you're told, you'll never get far in science. You have to prepare yourself for your future career. Researchers often get three or four years into their PhD before they even think about what's next.

You have to show initiative. Talk to students and postdoctoral researchers about holding career seminars in your department. Get involved with committees. Organize an event. Help to find sponsors. You'll develop a network of contacts that will greatly expand your options for the future. And you'll learn a lot about yourself. If you've never managed a budget, you'll find out whether you could see yourself doing that for the rest of your career.

It helps if your supervisor supports these activities. If not, try going higher up the ladder. Approach your department head or the agency that's providing your funding. Often, people at these levels are more supportive of career development.

These sorts of things take time, but it's all relative. Just think about how much time you spend in the lab working on things that don't pan out. You can spare a few hours for a meeting without sacrificing your science. If you're stuck in a place that doesn't value work outside the lab, you can do your part to change that culture.

Kruger wrote the article 'Why it is not a 'failure' to leave academia' (*Nature* **560**, 133–134; 2018).

IRINI TOPALIDOU Know yourself

Molecular biologist at the University of Washington in Seattle

Scientists often see a linear career pathway, and they think that's the only way to go. I see so many young principal investigators who are unhappy. They feel like they were pushed into their position, like they didn't have a choice. But if you know what you want to do and what you're good at, you can find a niche that's right for you.

If you're not sure whether you really like working at the bench, try it for three months to see whether it's right for you. If you aren't really excited about lab work, you can become easily

derailed by failure. You can also learn a lot about yourself by talking to others. Find someone who will listen to your concerns and offer advice without forcing you down a particular path.

And before you think about running your own lab, you should ask yourself a crucially important question: do you really have what it takes to be a good mentor? If you're not good at training people, or if you care more about your experiments than about your team, you probably shouldn't be a mentor unless you can dramatically change your approach. If more scientists were more self-aware, there would be fewer bad mentors, and the whole system would be better off.

RESEARCH SCIENTIST: ANOTHER WAY TO LEAD

If you like the idea of running a lab without dealing with a million administrative duties, give some thought to becoming a research scientist. It's not a fall-back position for people who can't make it as professors. Far from it. I'm very ambitious. But I didn't want to be the person who sits behind the closed door. As a research scientist, I can be a leader in the laboratory.

Some research-scientist positions are more rewarding — and more secure — than others. Find a lab that really needs you. Perhaps one with a less-experienced principal investigator who needs help building a lab. Or one helmed by a senior person who is too busy going to conferences to handle the day-to-day needs of a lab. If you make yourself valuable, you can expect to be valued.

Topalidou wrote the article 'Teach undergraduates that doing a PhD will require them to embrace failure' (*Nature* <https://doi.org/10.1038/d41586-018-06905-0>; 2018).

MIRJANA POVIC Connect to the developing world

Astrophysicist at the Ethiopian Space Science and Technology Institute in Addis Ababa.

More scientists should consider sharing their experience and knowledge in developing countries. Their expertise can go a long way in Ethiopia and many other countries in Africa, Asia and South America. But the benefits flow both ways. You can make huge personal and professional progress by going outside your normal routine and comfort zone. You learn many things when you adapt to different conditions, and supervising master's and PhD students with totally different perspectives can help you to tackle problems from new angles.

I moved from Europe to work in Africa, far from my home country of Serbia. A lot of Africans are doing great science, and they welcome collaboration. I have many colleagues

who work in Europe but come to Africa to give classes. Some supervise students or give lectures remotely. There are many ways to contribute.

Wondering how to get started? Try contacting a researcher in your field who is already working in a developing country. I'm always happy to share advice and information about different opportunities for research and collaboration. Astronomers can also contact the International Astronomical Union. Other branches of science have similar organizations that can point people in the right direction.

This life isn't easy. We have power outages and intermittent Internet. Sometimes it takes days to download data. But scientists can adapt and find ways to get things done. We learn new ways to do things and discover patience that we didn't know we had. That comes in handy in many areas of life.

Povic was the inaugural winner of Nature Research's Inspiring Science Award, developed in partnership with The Estée Lauder Companies (*Nature* **563**, 148; 2018).

ANDY KAH PING TAY Reach out for help

Biomedical engineer at Stanford University in California.

I learned as a PhD student that you can save a lot of the time and hassle of troubleshooting by avoiding the trouble in the first place. When you're trying a new technique, so many things can go wrong. A protocol might miss out certain key details. For example, how do you arrange the microscope? How long do you let a sample set before you do your imaging? And can you save money with a cheaper reagent?

Instead of just hoping that I understand a protocol, I get proactive. I e-mail researchers who have published papers about the technique. I don't contact the corresponding author; usually that person is pretty busy. Instead, I ask the first author, who is usually more junior.

Reaching out to other experts seems like an obvious step, but a lot of people are surprised that I do it. They warn me that someone might try to steal my ideas. But I don't worry about that. Other researchers are often happy to hear that someone wants to replicate their technique and validate their work. But you can't verify their results if you're making mistakes along the way. ■

Tay contributes to nature.com/careers, where readers can share their experiences and advice. Guest posts are encouraged. Contact naturecareerseditor@nature.com for more information.

INTERVIEWS BY CHRIS WOOLSTON

These interviews have been edited for clarity and length.

SOLSTICE

An unusual invitation.

BY JOHN GILBEY

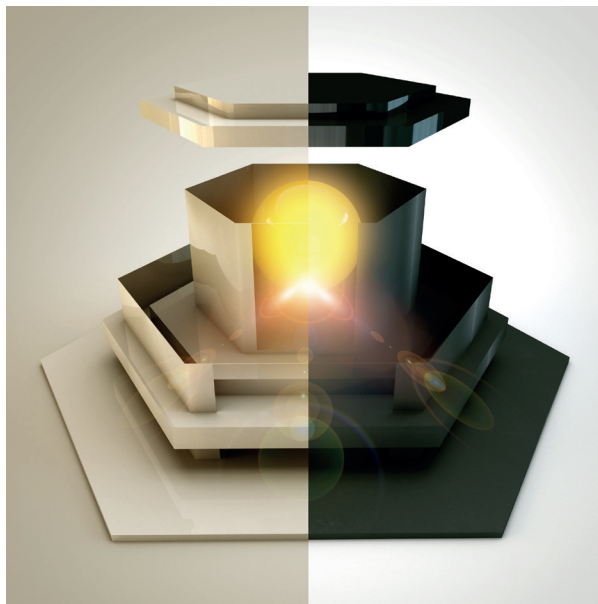
The letter on my desk wasn't from Human Resources, which was a huge relief, but from the Office of the Vice-Chancellor — inviting me to "The University of Rural England's Traditional Celebration of the Winter Solstice". I'd never heard of it, so assumed it must be one of those subtle traditions known only to the cognoscenti. Still, I was intrigued enough to e-mail my RSVP, pin the invitation to the shelf above my desk and blank out the evening of Friday 21 December on my calendar.

In my borrowed dinner suit, which smelt vaguely of mildew and old wardrobes, I presented myself at the anteroom of the Council Chamber just after 10 p.m.. The invitation was worryingly punctilious about timing — pointing out that "You will be admitted at 22:23 precisely", which a moment of study revealed was the local time of the solstice. At 22:22 the Head Porter appeared in his bowler-hatted splendour and unhitched the silken rope that barred access to the historic chamber. Then, as the clock ticked over to 22:23, he opened the double doors with a flourish and stepped aside.

The room was large, hexagonal — or maybe octagonal, it was difficult to judge — and walled with a mixture of oak panelling and huge mirrors. A capable team of hospitality folk circulated silently with trays of drinks and nibbles, both of which I took full advantage of. I wandered around munching, sipping and enjoying the music of the talented string quartet that was half-concealed behind a bank of festive shrubbery. The piece they were playing seemed familiar, but I couldn't quite place it — so I sneaked a look at the cover of the cellist's part and made a note to get a copy.

As time passed, the crowd seemed to swell — or maybe it was just folk getting louder as the booze took effect. I was wondering whether to look for the door and make my escape when an old gentleman, bald and bearded, stopped beside me. "John, isn't it? I wondered if I might see you here. I'd very much like to discuss your work, if you have a moment." I decided that I did, and the next hour passed

in a blur of conversation as I poured out to my new friend all the torment of my



current research failures and laboratory catastrophes.

He nodded, enquired, suggested solutions and alternative approaches until I began to wonder why I hadn't met him before — and why he wasn't my supervisor. Then it was over, the doors were opened again and the staff were easing a cheerful, very relaxed group of academics back out of the room — a million reflections echoing our departure.

I deserved the hangover I woke up with the following morning, but after a few extra coffees I started on the changes to my project that the old gentleman had suggested during our conversation. At lunchtime I trotted over to the library to find a copy of the quartet music, but returned frustrated, confused yet oddly fascinated.

Roll forward a year and I'd ironed out the major problems with the protocol, and the pilot hardware would be ready for beamtime in the spring — but there were still important things I wanted to discuss with my mysterious mentor. Without any contact details I'd failed to track him down and assumed he must be some visiting dignitary — so I was delighted when I got my second invitation to the solstice party.

The ceremonials were the same, and we filed into the Council Chamber a little before four in the afternoon. This time I found the old gentleman almost straight away: he was watching the string quartet and seemed pleased to see me. He asked about my work, but I pointed to the players — and asked him if he knew the music. He smiled, and

looked into his glass for a moment. I pressed on.

"He only wrote six quartets, yet this claims to be number seven." He looked up at me, evaluating my expression. "Ah, a subtle point, but illustrative nonetheless — and this is your field, after all ..." He steered me round to look at the room, the other guests, the chandeliers reflected in the mirrors.

"Have you ever noticed that there are those people you only meet at parties? Yes? Well, perhaps you could say this is just an extreme example of that phenomenon. This chamber, how many entrances would you say it has?" With the doors I'd entered through shut, they were indistinguishable from the panelling — it was impossible to say.

"Each guest has only one door available to them, but there are many more here than might be assumed.

Perhaps it was the monks who first built this chamber that grasped this, or maybe it was other, older folk. The university has certainly understood the special nature of this nexus for several hundred years — and used it to good effect."

When I asked him which door he had arrived through he shook his head. "It's difficult for me to judge — it may just be the one you used, although that is vanishingly unlikely. A myriad streams flow through this spot, yet by some quirk of celestial — or quantum — mechanics, they cross only at the moment of the winter solstice. Some guests, like us, find that we can talk to those from other realities — whose experience is subtly different, like the music of the quartet. A few of us have elected to remain within this nexus, never returning to our own space — to collate and distribute the benefits of truly parallel thought. Which reminds me, there is something I'd like to ask you ..."

I tried to picture the infinite intellects spiralling through the virtual space of the chamber, exchanging their most profound thoughts with the curators of this place. I took a glass of red wine from a passing tray and, holding out my free hand to the old gentleman, I smiled for the first time in months.

"I accept ..." ■

John Gilbey writes from the academic seclusion of the University of Rural England, where they worry an awful lot about this sort of thing. He is still waiting for his invitation, and tweets as @John_Gilbey.

ILLUSTRATION BY JACEY